Comprehensive Approaches to Identifying the Targets of Natural and Synthetic Antibodies Using Microarray DNA Synthesis and High-Throughput Sequencing

Comprehensive approaches to identifying the targets of natural and synthetic antibodies using microarray DNA synthesis and high-throughput sequencing

A dissertation presented by

George Jing Xu

to

the Committee on Higher Degrees in Biophysics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosphy

in the subject of

Biophysics

Harvard University

Cambridge, Massachusetts

September, 2015

Dissertation Advisor: Professor Stephen J. Elledge Author: George Jing Xu

Comprehensive approaches to identifying the targets of natural and synthetic antibodies using microarray DNA synthesis and high-throughput sequencing

Abstract

The incredible flexibility and specificity of the humoral immune response is dependent on the highly diverse repertoire of naïve and affinity-matured antibodies. Utilizing and understanding the power of this response requires high-throughput approaches. This thesis describes three projects that use recent advances in DNA sequencing and synthesis to develop and apply methods to probe the diversity of these responses at unprecedented depth. Chapter 2 describes a synthetic antibody library designed for high-throughput sequencing assisted selection which enables rapid *in vitro* selection of antibodies that bind specifically to a target of interest by bypassing the need for laborious single-clone screening for specific binding. Chapter 3 describes a high-throughput assay for detection of antibodies against all known human viruses using immunoprecipitation and high-throughput sequencing of bacteriophage displaying a library of peptides tiling through the proteome of all known human viruses. And last, chapter 4 describes the use of immunoprecipitation and high-throughput sequencing of both bacteriophage displayed peptides from the human peptidome and ribosome displayed proteins from the human proteome to identify a novel subclass of patients with scleroderma with autoantibodies against the minor spliceosome complex. The work described in this thesis will enhance our ability to study and exploit the properties of antibodies and the humoral immune response.

Table of Contents

Acknowledgements

Throughout my scientific education and training, I have received a great deal of help and positive influence from a whole host of incredible people. Here, I will attempt to personally acknowledge as many as possible. I am sure I have missed others who have had a significant impact on me, and to them I offer my sincerest apologies. Your kindness and generosity are truly appreciated despite my being remiss in these acknowledgements.

I would first like to thank all my primary and secondary school science teachers. I would especially like to thank Mr. Paul Ricks at Hopkins Junior High for constantly challenging us students to think deeply and teaching us the incredibly useful skill of searching the literature online. In addition, I would like to thank Mrs. Lisa Ishimine, Mr. Jim Camacho, Mr. Peter Geschke, Mr. Norman Howell, Mr. Jim Payette, and Mrs. Nathania Chaney-Aiello at Mission San Jose High School for putting together an incredible science curriculum that sparked my interest and served as a strong foundation for all my scientific endeavors. And extra thanks to Mr. Howell, Mr. Payette, and Mrs. Chaney-Aiello for agreeing to spend their precious free time chaperoning our unruly band of high school students to Science Bowl and Ocean Sciences Bowl competitions. I still have fond memories of studying and competing there and it would never have been possible without your generous support. I would also like to thank Professor Ken Shackel at UC Davis for supervising my summer research in his lab, as well as inviting me to his house and sharing his hobbies of windsurfing and didgeridoos, forever dispelling in my mind the myth of academic reseachers as solely bookish squares. And, of course, I have to thank Dave Varellas at the Young Scholars Program at UC Davis for being an incredible counselor during

the program and for letting me stay at his apartment after it ended in order to complete my research project. It was an amazing opportunity for which I will always be grateful.

I would also like to thank all of the great research mentors I had at Harvard College. First, I would like to thank Dr. Amy Rowat and Professor Dave Weitz for supervising my first college research experience during my freshman term. I would also like to acknowledge Drs. Tamara Brenner, Bill Senapedis, Nicholas Guido, Mike Strong, and Harris Wang and Professors Debra Auguste, Pam Silver, Jagesh Shah, William Shih, George Church, and Alain Viel for their time and dedication toward supervising our iGEM team's research. I would also like to give special thanks to Drs. Harris Wang and George Church for the three years of supervision and mentorship they gave me as an undergraduate researcher following the conclusion of iGEM. I would especially like to thank Harris for his close personal mentorship and guidance as I made my first foray into participating in a years-long research project. I learned an incredible amount about perseverance and creativity in scientific research in those three years – lessons that have served me well in my graduate studies.

In graduate school, I owe a special debt of gratitude to Dr. Ben Larman, who served as my rotation mentor and close research colleague during the early part of my graduate research. His excitement was infectious and his industriousness served as an incredible example. I would also like to give special thanks to Tomasz Kula, with whom I worked very closely during the latter half of my graduate studies. Despite supposedly being my junior, I always found his comments to be particularly insightful and discussions with him were always highly fruitful. I also want to thank Mamie Li, who would always kindly offer her excellent molecular biology advice whenever asked. And of course, I would like to thank Professor Steve Elledge himself for

the incredible opportunity to perform my graduate research in his lab, during which I had the great pleasure to benefit from the truly collaborative and exciting environment he has created with his group as well as learn an incredible array of skills for conducting rigorous and pioneering scientific research.

I would also like to thank my family, including my brother, David, and especially my parents, Felix and Chenny. They have supported me unwaveringly in all my endeavors and have made many sacrifices to afford me all the advantages and opportunities I am very grateful to have benefited from and, without which, I likely would not be here today.

Finally, I would like to thank the lovely Aileen Li, whose loving companionship brightened the frequent emotional lows of graduate school and without whom even the occasional highs would be savorless.

Chapter 1:
Introduction

Immunity depends on molecular recognition by specialized proteins

The human immune system is composed of a complex series of processes mediated by highly specialized cells, soluble factors, and physiological compartments. It is tasked with defending the body from invasion by both external threats, such as microorganisms and parasites, and internal threats, such as malignancies. To achieve this purpose, the immune system has a variety of powerful mechanisms for selectively controlling, killing, and eliminating invaders. These mechanisms must be tightly controlled to prevent unintended damage.

The basis of this control is discrimination between "self" and "non-self". "Self" encompasses all of the normal, healthy components that make up the host, whereas "non-self" can be almost anything else. As I will discuss further, the adaptive immune system is capable of recognizing an astounding diversity of "non-self" targets with exquisite precision. By limiting its destructive processes to "non-self" targets, the immune system can eliminate threats without damaging the body.

At the molecular level, recognition of "non-self" is mediated by proteins that bind specifically to molecular structures found only on "non-self" targets and not on the "self" (*1*). Both the innate and the adaptive branches of the immune system contain a variety of these recognition proteins.

The innate immune system uses germ-line encoded soluble proteins, such as mannose-binding lectins, and cell membrane associated pattern recognition receptors, such as Toll-like receptors, that recognize conserved pathogen associated molecular patterns and damage associated molecular patterns that are not normally found on undamaged "self" tissues (*2*).

Diversity of antibodies enables flexible recognition

The adaptive immune response, as its name suggests, is able to adapt the proteins it uses for recognition to novel molecular patterns that no germline-encoded protein is able to recognize. This remarkable flexibility arises from somatic recombination and mutation of a set of modular immunoglobulin gene segments that come together to form an incredibly diverse set of recognition proteins (*3*). In general, each round of recombination and/or mutation results in a clonal cell lineage that expresses a particular adaptive recognition protein with a unique specificity. The cells that produce these proteins can either be B cells, which secrete soluble antibody proteins that recognize extracellular targets (antigens), or T cells, which recognizes complexes of a specific peptide bound to a specific major histocompatibility complex (MHC) protein.

In this dissertation, I will focus on the humoral immune response, which is mediated by antibodies secreted by B cells. At any one time, each person harbors over 10^{11} B cells (*4*), each of which could be expressing an antibody from the over 10^{13} different theoretically achievable specificities (*5*). Maintaining such a diverse antibody repertoire increases the probability that at least one of them will be able to recognize any newly encountered antigen. Indeed, B cells can generate antibody responses against everything from small molecules, to proteins, to polysaccharides (*6*).

High-throughput sequencing enables study of antibody diversity

However, the incredible repertoire diversity that makes the antibody response so powerful also makes it extremely difficult to study. Traditional techniques for studying antibodies and their specificity, including hybridoma cloning, ELISA assays, and Sanger sequencing of antibody genes, are generally limited to assaying single or at most hundreds of

antibody-antigen interactions. With these technologies, it was simply not feasible to study the antibody response at any appreciable complexity.

In the past two decades, rapid advances in high-throughput DNA sequencing technology have led to novel methods for interrogating the diversity of antibody gene sequences at unprecedented depth (*7*). Since 1990, the per-nucleotide cost of DNA sequencing has decreased by almost seven orders of magnitude (*8*). This precipitous drop in price enabled several studies that were previously infeasible due to cost considerations. The results of these studies have broadened our understanding of the development of the humoral immune response and its function in a variety of settings, including infectious disease, cancer, autoimmunity, and immune deficiency (*7*).

High-throughput sequencing also improves selection of synthetic antibodies

These studies used high-throughput sequencing to examine the properties of pre-existing antibody responses. In Chapter 2, I discuss our work on using high-throughput sequencing to discover new antibodies that bind specifically to targets of interest. Because of their incredible specificity and affinity, antibodies have become indispensable tools for research, diagnostics, and therapeutics. In fact, monoclonal antibodies made up five of the top ten pharmaceutical products by sales volume in 2014 (*9*).

These antibodies are usually created either *in vivo*, by immunizing an animal with the target of interest, or *in vitro*, by using the target of interest to perform affinity purification of a library of synthetic antibodies (*10*). The *in vitro* method is generally simpler and faster because it

does not require animal husbandry or complex immunization protocols, but both methods result in a population of antibodies, of which only a portion will bind the target specifically.

Identifying the desired antibodies requires laborious screening of hybridomas or clonal species to ascertain their ability to specifically bind the target of interest. If the desired antibodies are not present at a significant proportion of the population, it may be necessary to screen an extremely large number of clones or perform additional immunizations or selections.

Because antibodies play such a critical role in research, diagnostics, and therapeutics, there is significant interest in simpler techniques to rapidly create antibodies against a range of targets. In particular, completion of the human genome sequence provided a catalogue of components of the human proteome, but the tools for studying these proteins are still lacking in comprehensiveness and quality. In response, several groups, including the Human Protein Atlas (*11*) and the Human Antibody Initiative (*12*), have sought to work collectively on developing a collection of monoclonal antibodies against each protein in the human proteome. Increasing the throughput of methods to create these antibodies would greatly enhance the ability to develop a comprehensive toolbox for proteomics research.

In our work, we develop an *in vitro* synthetic antibody library that was designed to use high-throughput sequencing to monitor enrichment (as a proxy for binding activity) of specific antibody clones in the entire population across multiple rounds of affinity purification to identify desired antibodies without the need for laborious clonal screening (*13*). The synthetic antibody library was designed using a common immunoglobulin framework and contained sequence diversity only in the relatively short complementarity determining regions so the identity of any member of the library could easily be obtained by sequencing these regions.

Since modern high-throughput sequencers can routinely sequence over 10^8 individual DNA molecules, this library design enables quantification of the relative abundance of each antibody in a very complex mixture. Antibodies that bind specifically are easily identified as the ones whose relative abundance increases after affinity purification on the target of interest but not a non-specific target. This analysis can be performed purely using the sequencing data and can detect antibodies at very low abundance without having to resort to laborious clonal screening. We demonstrated the application of this method to select a specific antibody against an antigen implicated in breast cancer cell transformation. The sequencing-assisted antibody library design and selection method we describe can be used to rapidly create antibodies that bind specifically to a target of interest and will enable more rapid development of tools for proteomic research, diagnostic assays, and therapeutic biologics.

Identification of an unknown antibody's antigen is important but difficult

Sequencing the genes encoding antibodies reveals the identity of the antibody, but it does not reveal the identity of the antigen the antibody recognizes. Despite recent progress in computationally modeling of the structure of immunoglobulins based on their amino acid sequence (*14*) and docking simulations to assess molecular interactions (*15*), predicting whether an antibody binds to a target based on sequence and structure alone are still extremely difficult and existing methods are usually inaccurate (*16*).

Thus, although high-throughput sequencing of the genes encoding antibodies has greatly increased the ability to study how individual antibody sequences arise and evolve, another method is required to study the antigen targets of these antibodies. The identity of the antigen targets often has greater clinical significance than the identity of the antibody. Multiple

antibodies with differing gene sequences could recognize the same or different regions on the same antigen, and knowing that any antibodies against an antigen exist can provide clinical insight into pathogen exposure and underlying disease processes. For example, if the antigen is a viral protein, the existence of antibodies against the antigen suggests prior infection. Or if the antigen is a self-protein, it could be suggestive of autoimmunity.

Because people are exposed to a diverse array of antigens in the environment and generate antibodies that could potentially target any number of those antigens, high-throughput technologies are required to determine which antigens the antibodies recognize. To test all of the possible antibody-antigen interactions, it is necessary to have a collection of all of the potential antigens of interest. Traditionally, in order to construct these collections in a cost-effective manner, researchers have turned to cDNA libraries of antigens and affinity purification using the unknown antibody or antibody mixture (*17*). However, the representation of each antigen in a cDNA library can vary wildly depending on its expression level and some antigens may not even be present if the protein is not expressed in the source RNA material (*18*). With such skewed libraries, it can be difficult or impossible to detect interaction between the antibody and an antigen that is expressed at very low or non-existent levels.

Advances in DNA synthesis enables technologies to identify antibody-antigen interactions

Instead of creating the antigen collection from natural material, recent advances in DNA synthesis have enabled the creation of fully synthetic representations of the antigen collection that are comprehensive and uniform (*19*). Reminiscent of the rapid decline in cost of DNA sequencing, the per-nucleotide price of oligonucleotide synthesis has dropped by over 7 orders of magnitude in the past three decades (*8*). These rapid advances in oligonucleotide synthesis have

largely been driven by the development of DNA microarray technologies that apply micro- and nanofabrication techniques to solid-phase nucleotide synthesis chemistries. Modern iterations of these technologies can synthesize over 10^5 custom oligonucleotides of up to hundreds of nucleotides in length at a price affordable to single researchers.

As was seen with the rapid advances in high-throughput sequencing, breakthroughs in DNA synthesis have also enabled a wide variety of new technologies that have transformed multiple fields of study (*20*). In this thesis, I will focus on their application in creating synthetic representations of protein antigens to study antibody-protein interactions.

In Chapter 3, I present work on the construction and validation of VirScan, an approach to comprehensively study the human antiviral antibody response using a library of viral protein fragments encompassing all known human viruses and its application in studying the antiviral antibody response across large populations. Viruses are traditionally considered purely in the context of pathogenesis, but recent studies have shown that human viruses are able to modulate host immunity by a variety of much more subtle mechanisms. In addition, vaccinations or productive infections can elicit long-lived plasma cells that can continue secreting their antibodies for decades (*21*). Thus, the immunological imprint on the circulating antibody repertoire left by the virus serves as a record of virus exposure and a window into the interactions between host immunity and viral infection.

Traditional methods to study the antiviral antibody repertoire, such as enzyme-linked immunosorbent assay (ELISAs), are generally limited to testing single or at most hundreds of antibody-virus or antibody-viral antigen interactions. Recently, viral metagenomic sequencing studies, enabled by the rapid advances in high-throughput DNA sequencing, have revealed that

humans are colonized or exposed to a highly complex collection of viruses (*22*). Studying the antibody response against all of these viruses and, in particular, all of the viral antigens are infeasible using standard techniques.

Instead, we took the approach of using the capabilities of microarray DNA synthesis and bacteriophage display to create a synthetic representation of protein fragments tiling through the proteomes of all human viruses that have been uncovered by DNA sequencing. Similar to the work in Chapter 2, we designed this library to be compatible with high-throughput sequencing, so that we could identify which viral antigens are targeted by a mixture of antibodies simply by using sequencing to find which antigens' relative abundances are increased after affinity purification with the antibodies. This approach enables comprehensive study of the humoral antiviral response in antibody samples from a large number of human donors. It incorporated many elements from Phage Immunoprecipitation-Sequencing (PhIP-Seq), a technology developed previously by our laboratory, and discussed in the following paragraphs (*19*).

In Chapter 4, I present the work on identifying a novel subclass of patients with the autoimmune disease scleroderma who have autoantibodies against the minor spliceosome complex. This work used both PhIP-Seq and a complementary approach, Parallel Analysis of Translated Open Reading Frames (PLATO) (*19*, *23*). Both technologies combine a collection of synthetic DNA encoded proteins and high-throughput sequencing to identify human autoantigens recognized by autoantibodies. PhIP-Seq is the predecessor of VirScan and pioneered the concept of combining high-throughput sequencing assisted affinity purification and a bacteriophage display library of protein fragments encoded by oligonucleotides synthesized on DNA microarrays. As such, it also has the advantages that VirScan has over traditional antibody-

antigen assays. However, because it uses protein fragments rather than full-length proteins, PhIP-Seq may miss antibodies that recognize discontinuous epitopes that cannot be capture by a protein fragment. PLATO overcomes this limitation by using ribosome display of full-length proteins, but its coverage of the human proteome is incomplete at present. Thus, combining these complementary technologies provides a powerful approach to identifying autoantigens.

Scleroderma is an autoimmune rheumatic disease of unknown etiology (*24*). Traditional serum ELISA and immunoprecipitation-Western blot assay have identified three mutually exclusive subclasses of patients with scleroderma who have autoantibodies against the centromere, topoisomerase 1, or RNA polymerase III (Pol III) (*25*). Intriguingly, patients in the subclass with autoantibodies against Pol III are also frequently diagnosed with coincident cancer and a recent study suggests the anti-Pol III autoantibodies may be due to a cross-reactive immune response against a tumor neoantigen in the gene encoding Pol III (*26*, *27*).

A significant percentage of patients with scleroderma do not fall into these subclasses and the targets of their autoantibodies are largely unknown. Interestingly, many of these patients are also diagnosed with coincident cancer. (*27*) We hypothesized that there are novel subclasses of patients with shared autoantibodies that arise due to the same tumor-associated cross-reactivity as the Pol III subclass so we used PhIP-Seq and PLATO to search for these novel autoantigen subclasses.

Chapter 2:
Construction of a rationally designed antibody platform for sequencing-assisted selection

H. Benjamin Larman[a,b,c,*], George J. Xu[a,c,d,*], Natalya N. Pavlova[c] & Stephen J. Elledge[c]

a. Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA
b. Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
c. Department of Genetics, Harvard University Medical School, and Division of Genetics, Howard Hughes Medical Institute, Brigham and Women's Hospital, Boston, MA, USA
d. Biophysics Program, Graduate School of Arts and Sciences, Harvard University, Cambridge, MA, USA

*These authors contributed equally to this work.

Acknowledgements

We thank Uri Laserson for sharing IgG heavy-chain sequencing data, François Ehrenmann and Marie-Paule Lefranc for providing crystal structure data, and Andreas Plückthun for providing p4c11L34Ser. S.J.E. is an investigator with the Howard Hughes Medical Institute.

Author Contribution

S.J.E. conceived and supervised the project. H.B.L. designed and constructed the HMM scFv library and optimized the ribosome display protocol. G.J.X. performed the selections, analyzed the sequencing data, and performed candidate validation studies. N.N.P. provided the PVRL4

antigen and performed the FACS analyses. The manuscript was prepared by H.B.L. and edited by G.J.X., N.N.P. and S.J.E.

Adapted from Larman, H. B., G. Jing Xu, N. N. Pavlova, and S. J. Elledge. "Construction of a Rationally Designed Antibody Platform for Sequencing-assisted Selection." *Proceedings of the National Academy of Sciences* 109.45 (2012): 18523-8528.

Abstract

Antibody discovery platforms have become an important source of both therapeutic biomolecules and research reagents. Massively parallel DNA sequencing can be used to assist antibody selection by comprehensively monitoring libraries during selection, thus greatly expanding the power of these systems. We have therefore constructed a rationally designed, fully defined single-chain variable fragment (scFv) library and analysis platform optimized for analysis with short-read deep sequencing. Sequence-defined oligonucleotide libraries encoding three complementarity-determining regions (L3 from the light chain, H2 and H3 from the heavy chain) were synthesized on a programmable microarray and combinatorially cloned into a single scFv framework for molecular display. Our unique complementarity-determining region sequence design optimizes for protein binding by utilizing a hidden Markov model that was trained on all antibody-antigen cocrystal structures in the Protein Data Bank. The resultant $\sim 10^{12}$-member library was produced in ribosome-display format, and comprehensively analyzed over four rounds of antigen selections by multiplex paired-end Illumina sequencing. The hidden Markov model scFv library generated multiple binders against an emerging cancer antigen and is the basis for a next-generation antibody production platform.

Introduction

Antibodies are useful for their ability to bind molecular surfaces with high affinity and specificity. The genetic basis for their structural diversity is partially encoded in the germ line, but is also the result of stochastic genetic events, including chromosomal rearrangements, nontemplated nucleotide insertions, and somatic hypermutation. The majority of this diversity is localized to the complementarity-determining regions (CDRs), which are the six-peptide loops that protrude from the variable domain framework to form the antigen-combining surface of the

antibody molecule. Three CDR loops are contributed by the heavy chain (H1, H2, and H3) and three by the light chain (L1, L2, and L3). CDRs 1 and 2 are encoded in the germ line, and are thus more constrained in their diversity. L3 is characterized by "junctional diversity," formed during the recombination of two gene segments (V and J). Finally, H3 is formed by two consecutive genetic rearrangements (first between D and J, and then between V and DJ), and is additionally accompanied by nontemplated "N" nucleotides, making this CDR the source of most naturally occurring antibody diversity.

Our goal was to develop a synthetic antibody production platform inspired by nature, which could be seamlessly integrated with massively parallel, short-read DNA sequencing analysis (Figure 1*A*) (*28*, *29*). For maximum convenience, we required that library amplification and sequencing reactions should depend upon a single set of primers, rather than the complex mixture necessary for natural repertoire amplification and analysis. Like others before, we therefore constructed a highly diverse antibody library within a single variable-domain framework (*30*, *31*). However, because it is well known that the natural diversity of variable-domain frameworks contributes to a naive repertoire's functional shape space (*32*, *33*), we sought to maximize the functional diversity in our library's CDR repertoire by rationally designing sequences based on a mathematical model of antibody–antigen interaction.

A single-chain variable fragment (scFv) is the simplest functional representation of an antibody molecule, and has become the platform of choice for most antibody engineers. Our first step was thus to identify the most suitable scFv framework to house libraries of rationally designed CDRs. Lloyd et al. screened a very large preimmune human scFv library against a panel of 28 different antigens, and after sequencing >5,000 postselection clones, they observed

strong enrichment of a small subset of heavy- and light-chain variable domains (*34*). Among these domains, the most highly enriched were the heavy chain VH1–69 and the λ-light chain VL1–44. The authors attributed these framework enrichments to increased expression and optimal folding within the periplasm of the *Escherichia coli* host cells. These findings were further corroborated by the work of Glanville and colleagues (*35*). We therefore housed our CDR libraries within an scFv framework composed of VH1–69 and VL1–44.

As a source of inspiration for CDR design features, we looked to the international ImMunoGeneTics' (IMGT's) annotated database of all antibody–antigen cocrystal structures present within Protein Data Bank (IMGT/3Dstructure-DB) as of May 2009 (*36*, *37*). Amino acid residues within CDRs can contribute to antigen binding in two distinct ways: (*i*) direct, via contribution of a side group that makes contacts with the antigen, and (*ii*) indirect, affecting the conformation of the peptide backbone in a way that permits the direct interaction of neighboring amino acid side groups. This behavior of CDR amino acid sequences can be captured in a two-state hidden Markov model (HMM). The "contact" state should be enriched for amino acids capable of sharing/exchanging electrons or burying hydrophobic surfaces, whereas the "noncontact" state should be enriched for residues capable of appropriately constraining or relaxing the CDR polypeptide backbone. An important feature of HMMs is that the state of each position depends upon its nearest neighbor. It is thus important to note that traditional approaches to synthetic CDR construction typically use degenerate nucleotides or codons, and so cannot link the identity of a particular residue to that of its neighbors. To implement our HMM design, we took a different approach and synthesized complete CDR sequences as releasable oligonucleotides on a programmable DNA microarray. Importantly, this approach permits the filtering of deleterious sequences, such as restriction sites and undesirable peptide motifs (e.g.,

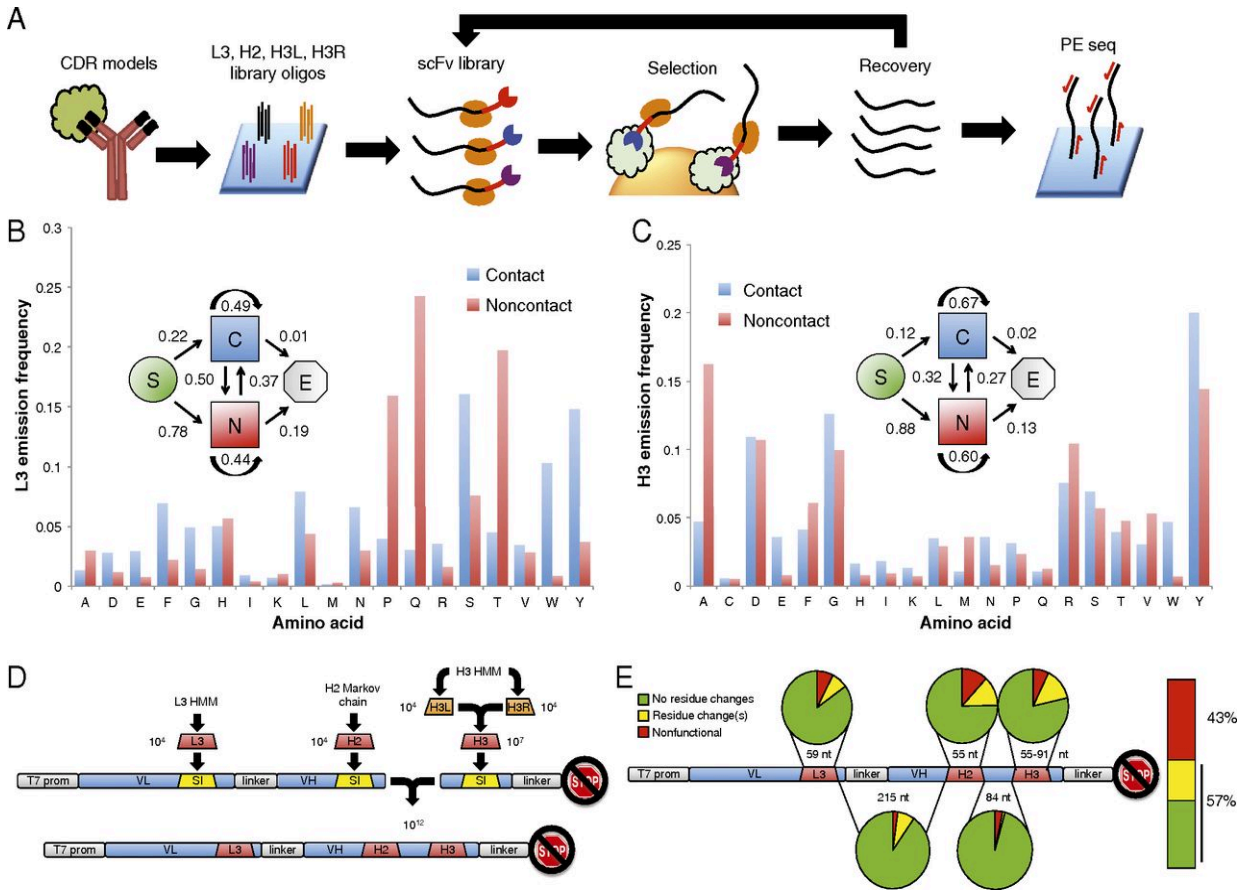

Figure 1. HMM antibody library design and synthesis. (*A*) Strategy for design and application of the rationally designed scFv library. Antigen–antibody crystal structures are used to design CDR-encoding DNA sequences, which are then synthesized on a programmable microarray. After ribosome display and enrichment for antigen binding clones, library recovery, and analysis by paired-end sequencing can be performed. (*B*) Model-defining parameters for the L3 HMM. Emission probability for each amino acid corresponding to the two possible states. State transition probabilities are inset: "S" denotes start of a chain, "C" denotes the contact state, "N" denotes the noncontact state, "E" denotes the end of the chain. (*C*) Model-defining parameters for the H3 HMM. Definitions are the same as for *B*. (*D*) Overview of the scFv ribosome display vector and library assembly strategy. "VL" and "VH" are the light and heavy variable domains, respectively. "T7 prom" is the T7 promoter, and the crossed stop sign denotes lack of a stop codon. L3, H2, and H3 are the CDR libraries designed to replace the "SI" suicide inserts. H3L and H3R sublibraries are brought together by combinatorial ligation to create H3. Similarly, the L3-H2 fragment is brought together with the H3 fragment in a combinatorial ligation. (*E*) Clonal Sanger sequencing analysis of 93 HMM scFv library members.

glycosylation signals and good HLA class II substrates), thus maximizing the functional utility of the library.

The transformation efficiency of bacterial cells with plasmid DNA is a significant barrier to construction of molecular libraries with a complexity greater than $\sim 10^{10}$. Because the utility of an scFv library scales with its diversity, we took advantage of the in vitro ribosome display technique, which has been used to generate antibodies with picomolar affinities (*38*). In this approach, mRNA molecules are tethered to the proteins they encode via noncovalent interactions with a ribosome. The mRNA is made to lack a stop codon necessary for peptide release, and so a population of ternary complexes composed of mRNA, encoded scFvs, and ribosomes is thus formed. Ribosome display libraries can be constructed and transcribed entirely in vitro, thus bypassing transformation bottlenecks.

After characterizing the quality of the HMM scFv library, we tested it by sequencing the library as it evolved over multiple rounds of selection on a protein antigen. We also developed robust methods to specifically recover desirable clones for expression and analysis in a simple two-step process. Our platform successfully produced antibodies against the emerging cancer antigen poliovirus receptor-related 4 (PVRL4) and sets the stage for a new paradigm in sequencing-assisted selection of rationally designed human antibodies.

Results

Library Design, Assembly, and Characterization

We set out to diversify the three CDR loops most relevant to antigen binding. By examining the IMGT/3Dstructure-DB, we determined the average number of contacts per

structure contributed by each CDR. Of contacts reported in this database, 76% are contributed by residues contained within CDRs. As expected, L3 and H3 contribute the most contacts, with H2 providing the third-most. In sum, 71% of CDR contacts are made by amino acids in these three CDRs (Figure 2A).

To estimate the HMM-defining parameters for L3 and H3, we identified 236 unique L3 and 241 unique H3 sequences within IMGT/3Dstructure-DB. Each residue was classified as either making contact or not with the protein antigen, as determined by the corresponding 3D cocrystal structure. The resulting HMM state transition rates and amino acid emission probabilities for L3 and H3 are illustrated in Figure 1B and C. Notable features of these models are: (*i*) enrichment for the noncontact state at positions closer to the framework [i.e., probability of S (start) → N (noncontact) and N → E (end) transitions are much greater than S → C (contact) and C → E, respectively]; (*ii*) in H3, a tendency for blocks of contact/noncontact states (i.e., probability of staying in the same state is higher than transitioning between states); (*iii*) a strong enrichment in both L3 and H3 for contacts consisting of tyrosine and tryptophan [previously observed by Ofran et al. (*39*)]; and (*iv*) L3- or H3-specific enrichments for certain amino acids in each state (e.g., noncontact proline in L3, and contact glutamic acid in H3).

We used our HMM to generate >10,000 unique sequences for each of L3 and H3 (*40*). Whereas the length of L3 sequences was fixed at 13 residues, 1,000 H3 sequences were randomly chosen for each length from 9 to 21 amino acids long. As an analog to VJ recombination, we further expanded the diversity of H3 by separating each sequence into two halves: “H3L” and “H3R,” for subsequent combinatorial ligation to form full-length H3 sequences (Figure 1D). This was accomplished by placing a type IIS restriction site downstream

of H3L and upstream of H3R on their encoding oligos. After PCR and restriction digest with the

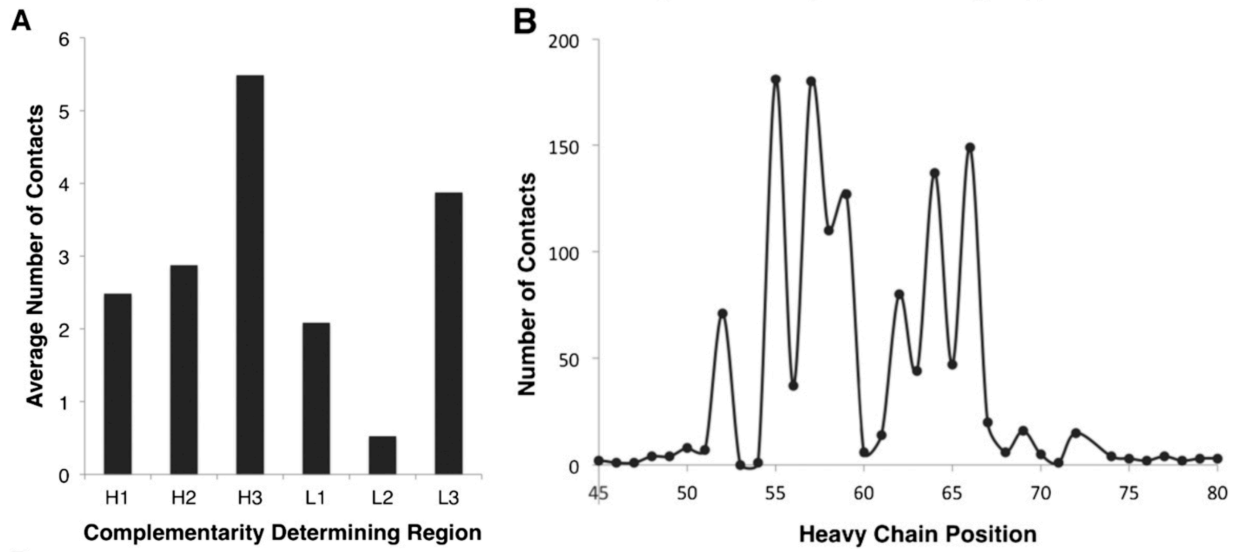

Figure 2. CDR contact distribution and H2 contact profile. (A) Contacts reported in the international ImMunoGeneTics/3Dstructure-DB database. Contact as- signment is based on international ImMunoGeneTics' definition of CDR positions. Data were obtained from 241 antibody-antigen cocrystal structures. (B) Position-dependent contact distribution in H2. Valleys represent amino acids more likely to play a role in framework stability.

SapI restriction endonuclease, the H3L and H3R fragments were combinatorially ligated together. The 3-nt 5′ SapI overhang on each sublibrary ensured that the H3 reading frames would remain intact.

The germ-line–encoded H2 CDR is characterized by structural features not present in L3 or H3 chains, and this is reflected in its heterogeneous contact profile (Figure 2B). It has been suggested that H2 contributes to the stability of the variable domain of the heavy chain through interactions among its hydrophobic residues (*41*, *42*). To avoid disrupting framework stability, we created a first-order Markov chain to generate H2 sequences that was trained on the 176 unique H2 chains in the IMGT database. This model was used to generate >10,000 H2 sequences.

Finally, all CDR sequences were passed through a series of three filters to maximize their utility. First, all restriction sites to be used during library construction were eliminated by introducing silent codon changes. Second, we sought to minimize the potential immunogenicity of the scFvs by discarding peptides with a high potential for loading onto HLA class II molecules during antigen presentation. We used the ProPred online server to filter our CDR sequences against the four most common HLA-DRB1 alleles (101, 301, 701, and 1501) with a stringency of 45% of the best substrate (*43*). This process resulted in replacement of about 18% of all H3 sequences by less immunogenic peptides. The third filter replaced sequences with the potential to interfere with industrial scale production (e.g., methionine oxidation, asparagine deamidation/cyclization), as well as glycosylation motifs.

The final set of 43,803 CDR sequences (L3, H2, H3L, H3R) were flanked by the appropriate restriction site sequences, as well as sublibrary-specific PCR primer binding sequences, and then synthesized as releasable oligonucleotides on a silicon wafer (Agilent Technologies). The oligo libraries were PCR-amplified and separately cloned into the VH1–69 and VL1–44 human heavy- and light-chain variable fragments for combinatorial assembly (Figure 1*D* and *Methods*). In vitro transcription was then performed to create the mRNA templates for ribosome display.

We characterized the HMM scFv library in two ways. First, we cloned a small sample of the library mRNA. This process allowed us to perform Sanger sequencing on individual colonies, and thereby estimate the overall fraction of the library expected to contain functional, full-length scFvs with no frameshift or nonsense mutations (57% functional, $n = 93$) (Figure 1*E*). None of the colonies examined had retained their CDR "suicide insert," and none had multiple

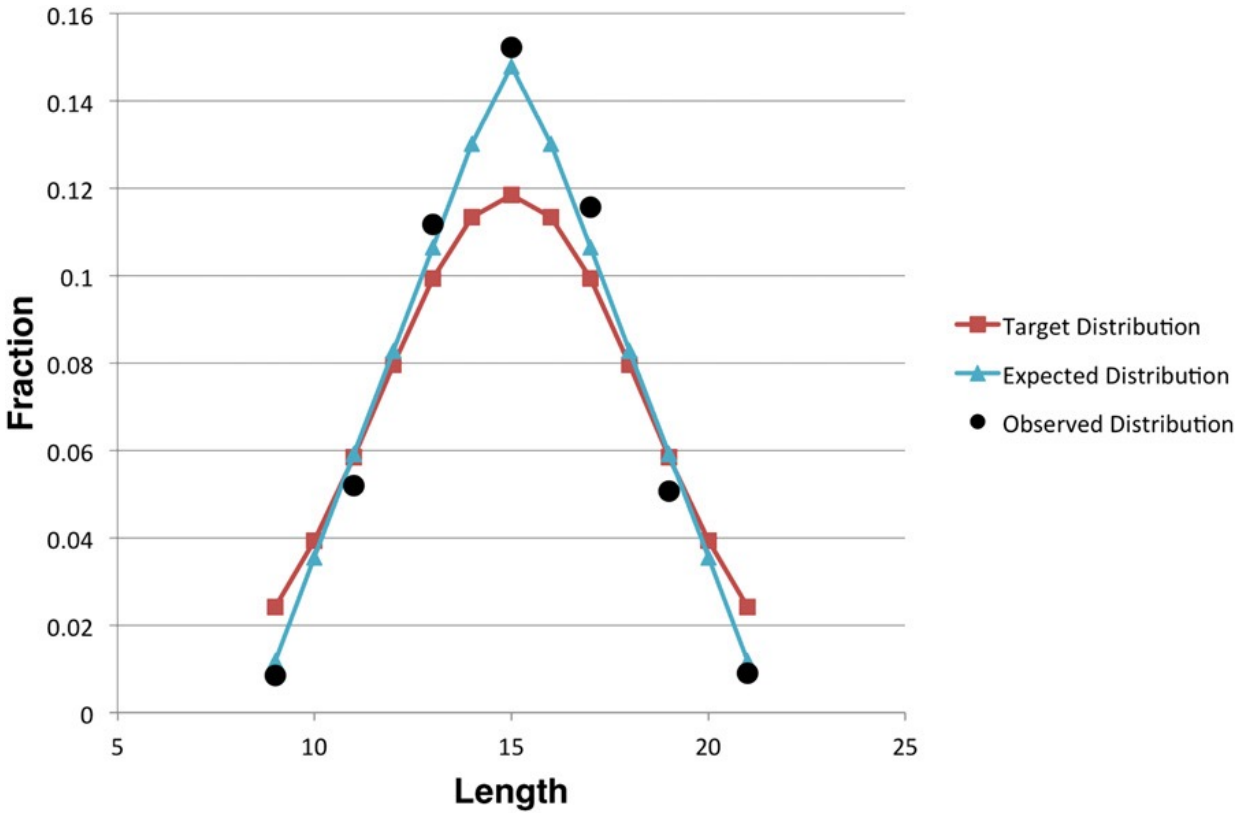

Figure 3. Length distribution of the H3 CDR library. Target H3 length distribution is based on the high-throughput sequencing of an individual's heavy-chain repertoire. Expected distribution is the calculated fraction of each length based on random ligation of all H3L sequences with all H3R sequences. The observed distribution is based on the analysis of the Illumina sequencing data from the unselected HMM scFv library.

CDR insertions. We found that the 43% nonfunctional clones derived mostly from oligo synthesis errors; ~15% of each CDR was nonfunctional, resulting in 39% (1–0.853) nonfunctional clones. Second, we used our Illumina sequencing data to characterize the length distribution of the H3 loop (Figure 3). Satisfied that our library was true to its design, we next performed selections against the cancer antigen PVRL4 (*44*, *45*), and used Illumina sequencing to track the library during selection.

Selection and Analysis of HMM scFv Libraries on GST-PVRL4 Bait

Four successive rounds of selection were performed with the HMM scFv library on GST-PVRL4. We quantified both enrichment of a spiked-in control scFv and the amount of RNA degradation during each round of selection to ensure these parameters met our quality-control thresholds (see *Methods*). After three selection rounds, the library was split and selected on either GST-PVRL4 or GST-GCN4 (no PVRL4) in parallel, which allowed us to discriminate between PVRL4-specific scFvs and those that bind to GST or to some other component of the system.

The minimal region of the HMM scFvs that contains the three diversified CDRs is an appropriate size for analysis by paired-end Illumina sequencing. The sequencing libraries can thus be prepared conveniently by PCR. A small amount of each of the selected libraries, as well as the unselected HMM scFv library was amplified with Illumina sequencing adapters. These adapters include a 7-nt barcode to permit multiplexed analysis.

We performed paired-end sequencing in two separate Illumina HiSeq 2000 flow cell lanes. By obtaining L3-H3 mate pairs in one lane and H2-H3 mate pairs in the other lane, we could use the hyperdiversity of H3 sequences to match corresponding L3 and H2 sequences, thereby reconstructing the complete identity of each scFv clone (Figure 4). After PCR amplification, however, we observed significant PCR chimerism, essentially resulting in CDR recombination. This complicated—but in most cases did not prevent—reconstruction of individual scFv clones. CDR recombination has been observed to significantly increase scFv affinity during ribosome display selection, suggesting that this process might actually improve the success rate of our platform (*46*).

We next determined the relative abundance of each clone in the library over the course of four rounds of selection on GST-PVRL4, and compared this to the results from the round 3

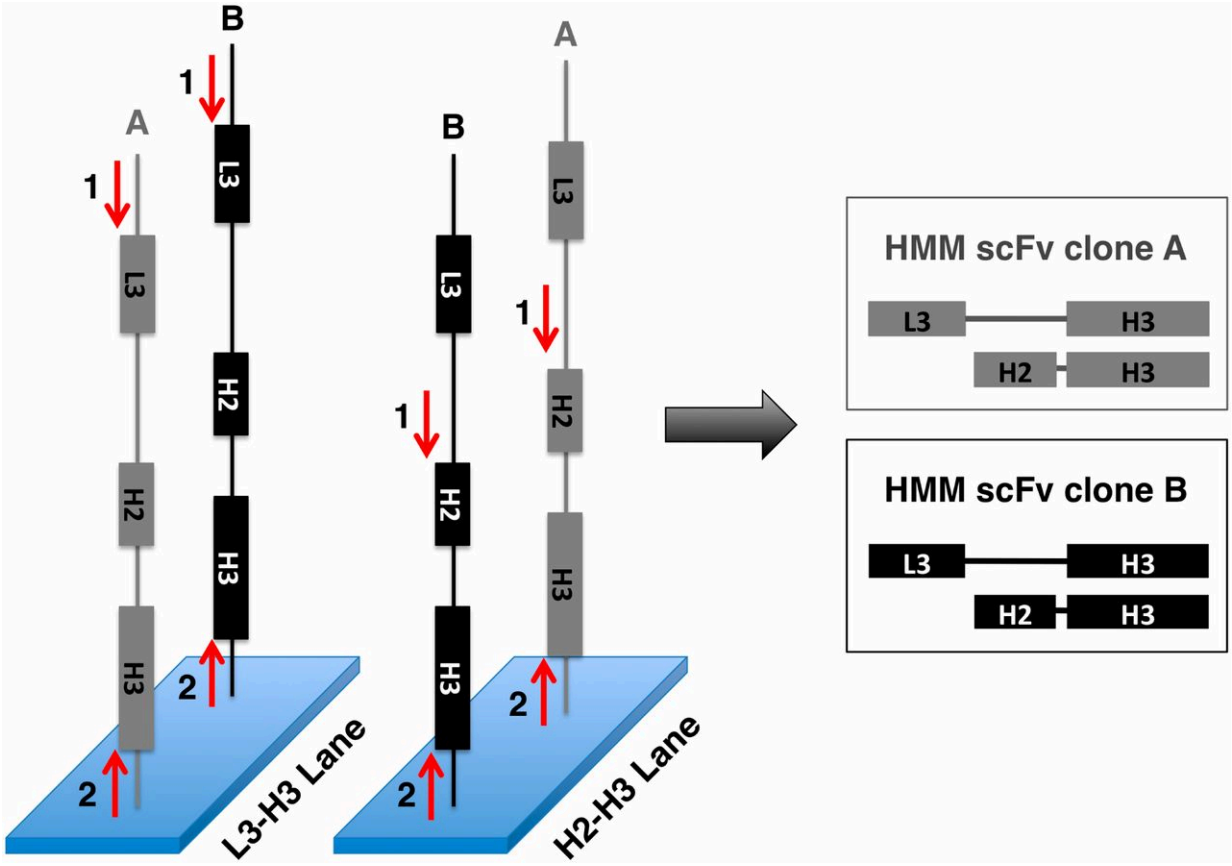

Figure 4. Strategy for sequence reconstruction of HMM scFv clones. One-hundred nucleotide paired-end sequencing is performed on the same library in two independent lanes on an Illumina HiSeq 2000. In the L3-H3 lane, the first sequencing primer lands upstream of L3 (red "1" arrow). In the H2-H3 lane, the first sequencing primer lands upstream of H2 (red "1" arrow). The H3 sequence is then determined by reading from a common, second primer (red "2" arrow) in both lanes. L3 and H2 sequences are then paired using their unique H3 identifier to fully define the sequence of the scFv clone.

PVRL4-selected library that was selected on GST-GCN4 (Figure 5). Based on this analysis, a subset of the candidate PVRL4-specific clones was selected for further analysis.

Recovery and Testing of Candidate Anti-PVRL4 HMM scFvs

Before characterizing individual scFvs for their ability to bind antigen, they must first be isolated. This isolation can be accomplished either by resynthesizing the CDRs for cloning back

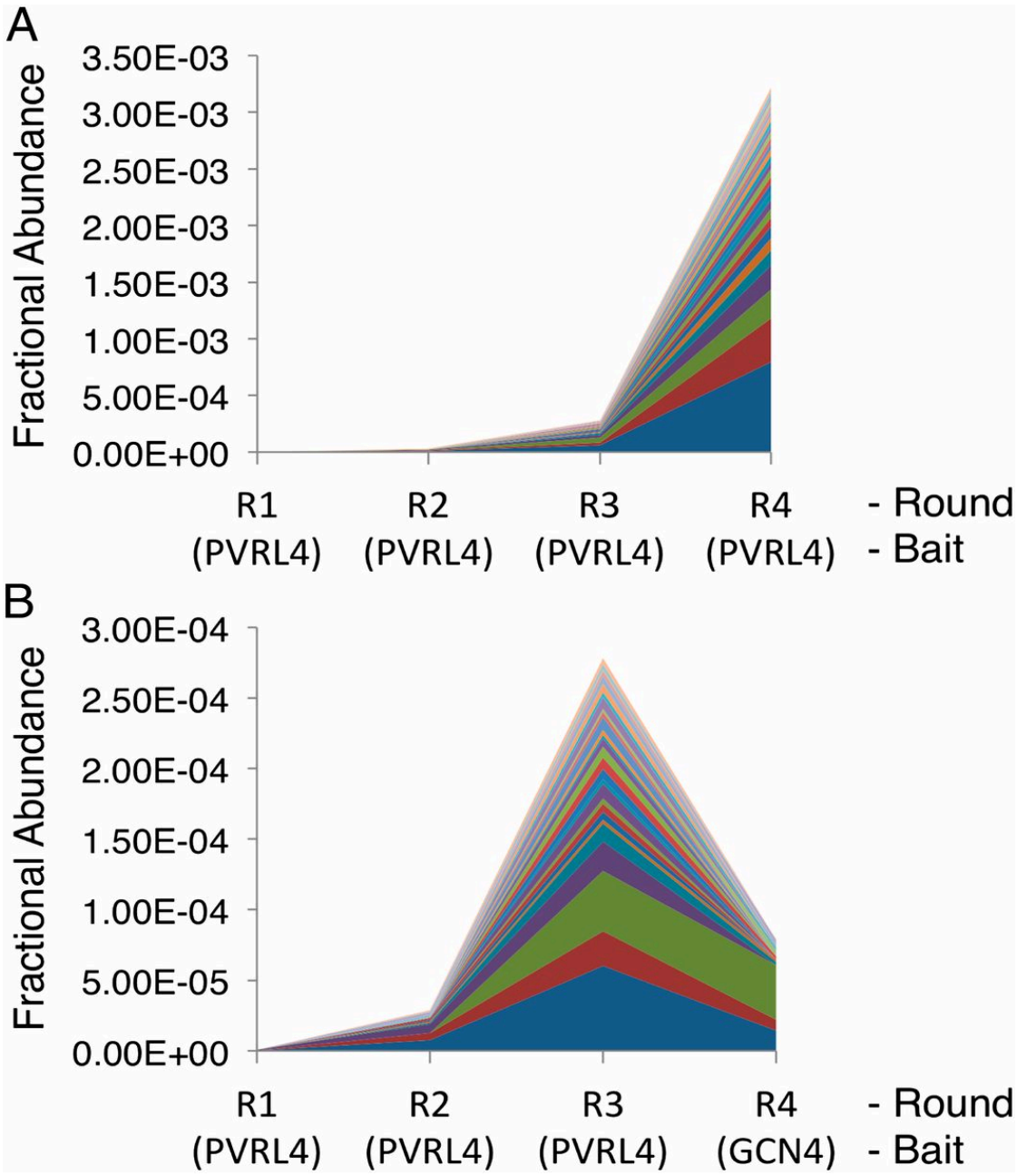

Figure 5. Fractional abundance of top 30 PVRL4-specific HMM scFv clones during selection. (*A*) The fractional abundance of the top 30 PVRL4-specific HMM scFv clones shown over four rounds of enrichment on GST-PVRL4. Fraction is calculated as read number of a clone divided by the total number of reads from the corresponding library. (*B*) The fractional abundance of the same 30 PVRL4-specific HMM scFv clones from *A*. Data from rounds R1–R3 are the same in the two panels. Round 4 selection on the non-PVRL4 bait, GST-GCN4, results in a relative depletion of these clones from the population.

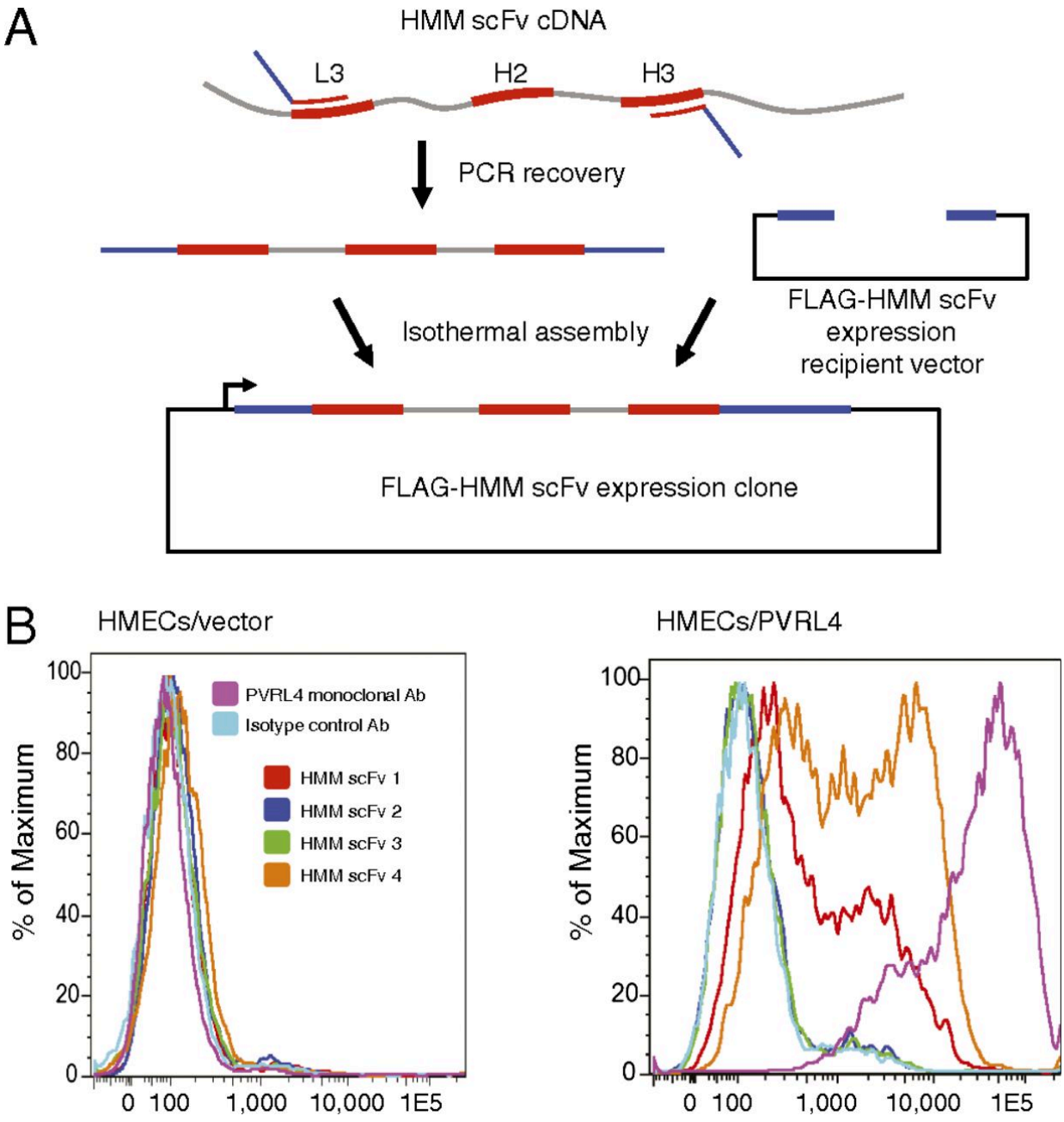

Figure 6. HMM scFv recovery strategy and FACS staining. (*A*) Candidate HMM scFv clones are recovered by PCR using primers specific for L3 and H3, and which also have 5′ homology arms for subsequent isothermal assembly into an expression vector with differing codon use. (*B*) Results of FACS staining to assess binding of candidate scFvs. HMECs infected with vector alone or vector-expressing PVRL4 antigen were stained with the indicated control antibodies or one of the top four most abundant postselection HMM scFv clones.

into an expression framework, or by PCR-recovering the clones using primers specific for L3 and H3. We chose to recover candidate scFvs by performing PCR with L3/H3-specific primers that also contained 5′ homology arms for subsequent isothermal assembly into a FLAG/6His epitope-tag expression vector (**Error! Reference source not found.***A*). Recovered candidate anti-PVRL4 scFv clones were expressed in vitro as FLAG-tagged proteins. Of the top 25 most abundant postselection clones, four were found to specifically bind human mammary epithelial cell (HMEC)-expressed PVRL4 by FACS-staining analysis (Figure 6*B*). The success rate of our method is likely underestimated by this analysis, however, as the selection bait was a bacterially-produced, unmodified GST-PVRL4 fusion, whereas HMECs display the glycosylated protein in the context of the cell surface.

Discussion

Synthetic biology has yet to deliver antibody production platforms that rival vertebrate immune systems in both product quality and manufacturing convenience (*47*). However, we anticipate that along with the maturation of gene-synthesis technologies and the affordability of high-throughput DNA sequencing will also come advances in antibody production pipelines that outperform animal immune systems in all regards (*48*).

To address the emerging need for a next-generation scFv production platform compatible with sequencing-assisted selection, we have created a rationally designed, single-framework scFv library. Single frameworks enable facile library amplification and sequencing using a single set of primers, but reduce framework-contributed functional diversity. To compensate for this potential loss of functionality, we used a mathematical model of structural data to capture subtle amino acid sequence biases that contribute to the formation of favorable antibody–antigen

contacts. We additionally developed a method to mimic the junctional diversity of VJ recombination using type IIS restriction cleavage followed by shuffling ligation. Our combinatorial strategy significantly reduces the required size of each CDR sublibrary, making extreme diversification of the final scFv library a tractable problem. Finally, to accomplish the analysis of selected scFv libraries using short-read DNA sequencing, we have introduced a strategy for double mate pair-based reconstruction of full-length scFv clones.

One benefit of the three-CDR library presented here is the ease of clonal sequence reconstruction. We found that the hyperdiversity of our H3 CDR library permitted the near unambiguous pairing of L3 and H2 sequences with their shared H3, thus completely defining the repertoire at each round of selection. We went on to demonstrate an important advantage of single-framework libraries, which is the relative simplicity of recovering desirable clones. Our two-step protocol enables rapid isolation and assembly of clones into an expression vector for further functional characterization. In contrast, native heavy- and light-chain framework libraries are currently intractable to analysis with short-read sequencing, and require many more steps for clonal isolation. Although not demonstrated here, an additional design feature built into our platform is the ability to perform straightforward total synthesis of desirable clones from synthetic DNA oligos or oligo pools (Figure 7). It is worth noting that the utility of the HMM library is not limited to the ribosome display format, and can be moved into traditional phage or yeast display vectors. Similarly, alternative heavy- and light-chain frameworks can be synthesized to house the HMM CDR libraries described in this work.

An unrealized application of sequencing-assisted selection is the parallel production of antibody sets that target multiple antigens. For example, an “array” of antigens can be pooled in

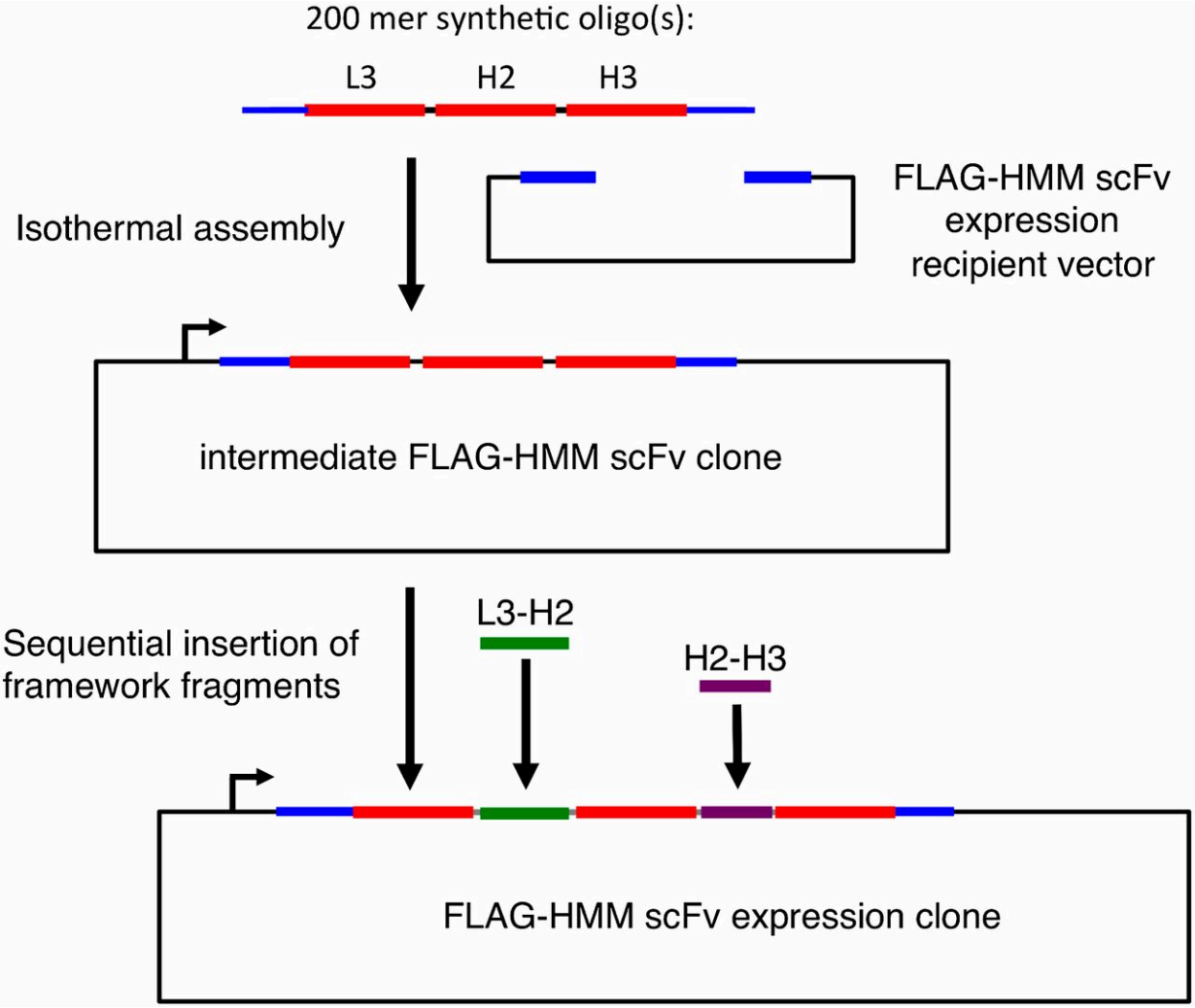

Figure 7. Total synthesis of HMM scFv clones. Strategy for reconstruction of desirable HMM scFv clones from synthetic oligos. After PCR amplification of the oligo or oligo library, the CDRs are assembled into an intermediate expression vector. The constant sequences of the framework regions between L3 and H2 (green) and between H2 and H3 (purple) are then sequentially inserted to form the complete HMM scFv expression clone or clone set.

different configurations, screened, and the resulting antibodies deconvoluted so that scFvs specific to a single antigen can be deduced (*49*, *50*). This strategy can reduce the number of selections to the square root of the number of antigens. Future single-pot, massively parallel selections will require the development of robust library-vs.-library deconvolution strategies, and preliminary progress has recently been reported (*51*).

As more sophisticated selection/deconvolution strategies emerge, and as immuno-PCR applications become more commonplace, we anticipate an increasing demand for efficient, low-cost production of high-quality synthetic antibody reagents. The methods presented here are a step toward this goal. As a first iteration, however, our platform is not without limitations. For example, not all of the scFvs predicted to bind PVRL4 were found to do so by FACS analysis. Indeed this finding could reflect the different antigenic context of cell surface PVRL4 or it could be the result of inaccurate scFv clone-enrichment quantification. Advances in DNA sequencing depth and read length will improve our ability to quantify clonal abundances, and will eliminate the need for the double mate-pair reconstruction of full-length sequences presented here. Read-length improvement will also facilitate analysis of libraries based on diversification of greater than three CDRs.

For this proof-of-principle work, we did not incorporate mutagenic PCR into the library-recovery protocol because it adds a layer of complexity to the analysis of enriched populations. However, the power of deep sequencing to map binding energy landscapes is now being realized (*52*) and will undoubtedly yield similar utility in the context of antibody selections. Finally, the profusion of additional antibody–antigen cocrystal structures will improve our ability to model CDR sequence biases that give rise to favorable antibody properties. With these considerations in mind, there is little doubt that sequencing-assisted selection of synthetic antibody libraries will play an increasingly important role in a wide variety of future biomolecular investigations.

Methods

HMM scFv Library Assembly

We wished to use the J chains most commonly associated with the VH1–69 and VL1–44 segments. In a sequenced heavy-chain repertoire from an individual, IGHJ4 was the J chain most often recombined with VH1–69. We used work by Schofield et al. to determine that in a large pool of selected phage, IGLJ2 was the J chain that most often recombined with VL1–44 (*53*). These components were assembled and reverse translated into an *E. coli* codon preference (Table 1). We introduced silent mutations into the framework regions flanking L3, H2, and H3, for the purpose of cloning in the CDR libraries. We required that at least one of each of these pairs be nonpalindromic so as to minimize multiple CDR insertions during library cloning. To this end, we introduced a BbsI site 5′ and an Acc65I site 3′ of L3, a PflMI site 5′ and an ApoI site 3′ of H2, an AccI site 5′ and a BstEII site 3′ of H3. These pairs of cloning sites flanked replaceable suicide inserts, which contain a stop codon in all reading frames and a XhoI restriction site. The CDR libraries were released from the microarray as 10 pmol of single-stranded DNA and resuspended in 200 μL water. Next, 1 μL of each sublibrary was used as input for library-specific PCR using 1 μL Taq polymerase (TaKaRa) according to the manufacturer's instructions (2 μM each primer). The thermal profile was: (*i*) 95 °C 5 m, [(*ii*) 94 °C 15 s, (*iii*) 55 °C 30 s, (*iv*) 68 °C 15 s] × 24. At this point, the reaction was divided in two and primers were replenished. The thermal profile was then: (*i*) 95 °C 5 m, (*ii*) 94 °C 45 s, (*iii*) 67 °C 7 m.

The L3 sublibrary was cloned into the scFv vector at the BbsI and Acc65I sites, electrotransformed into DH10B cells, and grown overnight on 15-cm carbenicillin plates. We harvested $>10^7$ transformants by scraping, and purified their plasmid DNA. Starting with this HMMscFv-L3 library, the same procedure was then used to replace the H2 suicide insert with the H2 library PCR product by using the engineered PflMI and ApoI sites. We obtained $>10^7$ transformants (HMMscFv-L3-H2 library) and purified the plasmid DNA. The H3L library PCR

Feature	AA Sequence	Nt Sequence	Remark
VL1-44	QSVLTQPPSASGTPGQRVTISCSGSSSNIGSNTVNWYQQLPGTAPKLLI YSNNQRPSGVPDRFSGSKSGTSASLAISGLQSEDEADYYC- L3_suicideInsert	CAATCTGTGCTGACCCAGCCACCGTCGGCCTCGGGTACTCCGGGTCAGCGTGT TACGATCTCCTGCAGCGGTTCTTCCTCTAACATCGGTAGCAACACGGTTAACT GGTATCAACAGCTGCCGGGCACTGCCCCAAAACTGCTGATCTACTCCAACAAC CAGCGTCCAAGCGGCGTTCCGGATCGTTTCAGCGGTAGCAAAAGCGGTACTTC CGCGTCCCTGGCGATCTCTGGCCTGCAGTCCGAAGACGAAGCGGATTATTATT GCtaataactcgagttaataactagttttaataaggtg	Framework in UPPERCASE, suicide insert in lower case
VL1-44_opt	QSVLTQPPSASGTPGQRVTISCSGSSSNIGSNTVNWYQQLPGTAPKLLI YSNNQRPSGVPDRFSGSKSGTSASLAISGLQSEDEADYYC- L3_suicideInsert	CAAAGCGTTCTGACCCAGCCTCCGTCCGCGAGCGGCACCCCGGGTCAGCGTGT TACCATTTCTTGTAGCGGTAGCAGCAGCAACATTGGTAGCAATACCGTCAATT GGTATCAGCAACTGCCGGGCACCGCACCGAAACTGTTGATCTACAGCAACAAC CAGCGCCCGAGCGGCGTCCCAGACCGTTTTCGGGCAGCAAATCCGGTACGAG CGCCAGCTTGGCGATCAGCGGTCTGCAAAGCGAAGACGAGGCCGATTACTACT GC taataactcgagttaataactagttttaataaggtg	Framework in UPPERCASE, suicide insert in lower case
IGVL-J3	FGGGTKLTVL	TTTGGCGGCGGTACCAAACTGACCGTTCTG	
IGVL-J3_opt	FGGGTKLTVL	TTCGGCGGTGGTACCAAGCTGACGGTGCTG	
VL-VH_Linker	GGGGGSGGGGSGGGGSSGGGS	GGCGGTGGTGGTGGCTCTGGTGGTGGGGGTTCCGGTGGTGGCGGCAGCTCCGG CGGTGGTTCC	
VL-VH_Linker_opt	GGGGGSGGGGSGGGGSSGGGS	GGCGGTGGTGGTGGCTCCGGTGGCGGTGGTTCCGGTGGTGGCGGTTCGAGCGG TGGCGGCAGC	
IGHV1-69	QVQLVQSGAEVKKPGSSVKVSCKASGGTFSSYAISWVRQAPGQGLE- H2_suicideInsert- YAQKFQGRVTITADEATSTAYMELSSLRSEDTAVYYC- H3_suicideInsert	CAGGTGCAGCTGGTGCAGTCCGGTGCGGAAGTTAAGAAACCGGGTTCCTCCGT AAAAGTCTCTTGCAAGGCGAGCGGTGGTACTTTCTCCTCCTACGCGATTTCTT GGGTGCGTCAGGCACCGGGCCAAGGTCTGGAAtgatgactcgagttgatgaga tatcttgatgagTATGCGCAGAAATTTCAGGGCCGCGTAACCATCACTGCCGA TGAGGCGACTTCCACCGCCTACATGGAGCTGTCTAGCCTGCGTTCTGAAGATA CCGCTGTCTACTACTGCtgataattaattaatgactcgagtttgataagg	Framework in UPPERCASE, suicide inserts in lower case
IGHV1-69_opt	QVQLVQSGAEVKKPGSSVKVSCKASGGTFSSYAISWVRQAPGQGLE- H2_suicideInsert- YAQKFQRVTITADEATSTAYMELSSLRSEDTAVYYC- H3_suicideInsert	CAAGTGCAGCTGGTGCAGAGCGGTGCAGAGGTTAAGAAACCGGGCTCTAGCGT AAAGGTGTCTTGTAAGGCCTCCGGTGGTACGTTCAGCAGCTATGCGATTAGCT GGGTTCGCCAAGCACCGGGCCAAGGCCTGGAAtgatgactcgagttgatgaga tatcttgatgagTATGCGCAGAAATTTCAACGTGTCACCATCACCGCTGACGA GGCTACTAGCACGGCGTACATGGAACTGAGCAGCCTGCGTTCTGAGGATACGG CGGTcTACTATTGCtgataattaattaatgactcgagtttgataagg	Framework in UPPERCASE, suicide inserts in lower case
VHJ3	WGQGTMVTVSS	TGGGGCCAGGGCACGATGGTGACCGTGAGCAGC	W is position 131 and varied according to composition
VHJ3_opt	WGQGTMVTVSS	TGGGGTCAGGGTACTATGGTGACCGTCAGCAGC	W is position 131 and varied according to composition
TolA	QKQAEEAAAKAAADAKAKAEADAKAAEEAAKKAAADAKKKAEAEAAKAA AEAQKKAEAAAAALKKKAEAAEAAAAEARKKAATE	CAGAAGCAAGCTGAAGAGGCGGCAGCGAAAGCAGCGGCAGATGCGAAAGCTAA GGCCGAAGCAGATGCTAAAGCTGCGGAAGAAGCAGCGAAAAAGGCGGCTGCAG ATGCAAAGAAGAAGGCAGAAGCAGAAGCCGCCAAAGCCGCAGCCGAAGCGCAG AAAAAAGCCGAGGCAGCCGCCGCGGCACTGAAAAAGAAGGCGGAAGCGGCAGA AGCAGCAGCAGCAGAAGCAAGAAAGAAAGCGGCAACTGAA	

Table 1

product was first NheI/BssHII subcloned into the pPAO2 vector (*54*). About 5×10^6 transformants were obtained and plasmid DNA collected. In parallel, H3R library PCR product was prepared. From the pPAO2-H3L plasmid pool, ~300 bp of upstream sequence was PCR-amplified for subsequent size discrimination of H3L-H3R ligation product. Both pPAO2-H3L and H3R PCR products were digested with SapI for subsequent combinatorial ligation by their 5′ overhanging codons. High-concentration T4 ligation was carried out at 15 °C overnight, a condition that permits mismatched ligation at a relatively high frequency. Indeed, upon sequencing a large number of H3 clones, we observed many examples of library members with unmatched codons that were ligated together, and importantly, without disrupting the reading frame. After H3 ligation, the correct size product was gel-purified and PCR-amplified. This PCR product and the HMM scFv vector were then digested with AccI and BstEII, so that the final H3 library could replace the vector's H3 suicide insert. If only complementary codons were able to

ligate together, the theoretical diversity of the H3 sublibrary would be 1.2×10^7. However, we frequently observed noncomplementary ligation, thus increasing the expected diversity of H3. About 107 H3 clones were obtained.

To bring together HMMscFv-L3-H2 and HMMscFv-H3 in a final ligation (Figure 1*D*), 60 μg of each of library was first digested with AccI and BbsI and the desired fragment gel-purified. In a high-concentration T4 ligation at 37 °C, the two fragments were ligated to form concatamers. Finally, the product was digested with both NotI (to release the desired in vitro transcription template) and XhoI (to destroy clones retaining a suicide insert) and gel-purified. We recovered 2.44 μg of HMMscFv-L3-H2-H3 library DNA at the correct size, which corresponds to 3.07 pmol or 1.85×10^{12}, theoretically unique DNA molecules. This material was used as a template for in vitro transcription (RiboMAX Large Scale RNA Production System T7; Promega) to produce mRNA, which was subsequently isolated with TRI reagent (Ambion).

Ribosome Display

Before immobilization of antigen-GST fusion protein, MagneGST beads (Promega) were washed 3× in 1× TBST. Five-microliter beads were used per immunoprecipitation, and beads were coated with 100 μL of bacterial lysate containing GST fusion protein mixed 1:1 with TBST. Two μL of 1M dTT were included. Binding occurred overnight by rotating at 4°C. RD Buffer, 1 L: 50 mM Tris Acetate (6.07 g), 150 mM NaCl (8.77 g), pH to 7.5 with acetic acid; autoclaved. Beads were washed 5× with buffer "RDWB+T" (RD Buffer plus 50 mM Mg Acetate and 0.5% Tween 20) and tubes were changed after every other wash. Beads were blocked in 50 μL "Selection Buffer" (RDWB+T plus 2.5 mg/mL heparin and 1% BSA and 83.3 μg/mL tRNA) plus 1 μL RNasin (Promega) at 4 °C for 2 h.

Next, 6.37 μg RNA (1×10^{13} RNA molecules) per 14 μ L translation reaction were used. Translations were performed using the RTS 100 *E. coli* Disulfide kit (5 PRIME) according the manufacturer's instructions, except that the feeding solution was not used. Translation was allowed to proceed for 13 min 45 s at 30 °C. Each 14 μ L reaction was immediately diluted with 96 μL ice-cold Selection Buffer and 3 μL RNasin. Reactions were centrifuged 14,000 × *g* for 5 min at 4 °C. Supernatant was then moved to a new, cold tube. Fifty-microliter beads in Selection Buffer was added to the ribosome-displayed HMM scFv library and rotated 4 h at 4°C. Beads were washed six times with 500 μL ice-cold RDWB+T. Tubes were changed after every other wash. Ribosomal complexes were disrupted after the final wash by resuspending beads in 50 μL "EB20" (RD Buffer plus 20 mM EDTA) plus 1 μL RNasin and incubated at 37°C for 10 min. Released RNA was then purified on Qiagen RNeasy column and eluted into 33 μL nuclease-free H2O.

Superscript III kit (Invitrogen) was used to reverse transcribe the selected RNA library from the preTolA primer. Next, 1 μL (5 U) of *E. coli* RNase H (New England Biolabs) was incubated with the RT product at 37°C for 20 min. Recovered cDNA was first PCR-amplified using primers that flank an insert region containing the CDRs (LLF2 and LLR2). PCR amplification was performed with the GC-RICH PCR kit (Roche) using the following the conditions: 1× GC-RICH Buffer, 0.2 mM of dNTP, 0.2 μM LLF2 primer, 0.2 μM of LLR2 primer, 0.5 μM of Resolution Solution, 1 μL of enzyme per 50 μL reaction. The thermal profile was: (*i*) 95°C for 3 min, [(*ii*) 95°C for 15 s, (*iii*) 55°C for 30 s, (*iv*) 72°C for 1 min] × 40 cycles, (*v*) 72°C for 7 min. The resulting PCR product was then double-digested with BbsI and BamHI (New England Biolabs), gel-extracted, and ligated using T4 Ligase into the pRDscFv2 vector. The ligation product was then PCR amplified using primers specific for the T7 promoter and the

TolA linker (T7B2 and TolA). PCR amplification was performed as above but with the T7B2 and TolA primers. The final PCR product was digested with XhoI (New England Biolabs) and gel-purified for either Illumina sequencing or use in a subsequent round of selection.

Recovery of HMM scFv Clones from a Selected Library

Single HMM scFv clones were recovered from the selected library by PCR with CDR-specific primers followed by assembly into a protein expression vector. Forward and reverse primers were designed to amplify target clone's L3-H2-H3 insert and contained a 20 bp adapter sequence for assembly into the protein expression vector. PCR amplification was performed with the following conditions: 1× Phusion High-Fidelity PCR Master Mix with HF Buffer, 0.2 μM each of the forward and reverse primers, 1 μL of cDNA recovered after library selection per 50-μL reaction. The thermal profile was: (*i*) 98°C for 30 s, [(*ii*) 98°C for 10 s, (*iii*) 72°C for 1 min] × 30 cycles, (*iv*) 72°C for 10 min.

PCR products were subsequently gel-purified and assembled into a protein expression vector using an isothermal assembly method. The protein expression vector contains the scFv framework followed by a FLAG tag and two in-frame stop codons. The isothermal assembly reaction was performed as previously described (*55*). Each reaction contained 100 ng of linear vector DNA lacking the L3-H2-H3 insert and 20 ng of the recovery PCR product, and was incubated at 50°C for 1 h. One-microliter of the assembly reaction product was transformed into DH5α *E. coli* cells and colonies were picked for sequence verification. Plasmids were expressed using the RTS 100 Disulfide Kit (5 PRIME) according the manufacturer's instructions, except that the feeding solution was not used. The resulting product was used directly in subsequent experiments.

Construction of the Ribosome Display Vector

Plückthun and colleagues have optimized vectors capable of displaying single-chain variable fragments (scFvs) on ribosomes (*56*, *57*). We adapted components of these and other such vectors to our present purpose. Beginning from the 5′ end of the DNA vector, the following parts were assembled as a synthetic gene product (DNA2.0):

1. T7 promoter for in vitro transcription from the DNA library (TAATACGACTCACTATAGGGAGACCACAACGGTTTCCC);

2. 5′ mRNA stemloop (5′-GGGAGACCACAACGGTTTCCC-3′) to improve transcript stability;

3. Ribosome binding site for translation of the library;

4. Kozak sequence for potential use in eukaryotic translation systems;

5. N-terminal 6xHis tag for detection and potential purification of scFv protein;

6. The variable domain of the light chain was encoded N-terminal to the heavy chain so that PCR recovery of the three diversified complementarity-determining regions (CDRs) (L3, H2, H3) would require the shortest amplicon;

7. Between the N-terminal variable light chain (VL) and C-terminal variable heavy chain (VH) is a "$(G4S)_3$" linker with optimized codon use (5′-ggtggtggtggttctggtggtggttctggcggcggcggctccagtggtggtggatcc-3′);

8. The C terminus of VH is fused to a linker segment derived from the TolA Escherichia coli protein (accession: NP_415267, position 131–214), which provides a spacer between the displayed scFv and the ribosomal tunnel;

9. 3′ mRNA stemloop (5′-CCGCACACCTTACTGGTGTGCGG-3′) to improve transcript stability.

NotI sites flank the 3′ and 5′ ends of the construct for isolation of the in vitro transcription template. Directional SfiI sites flank the minimal scFv for facile movement of clones into and out of alternative expression vectors.

Quality Control During scFv Selection

We used a positive control scFv and bait pair to optimize our ribosome display selection protocol. Pluckthun and colleagues have used ribosome display to affinity mature an scFv (4c11L34Ser, "Pluck-scFv") to high affinity (K_d = 40 pM) for a peptide derived from the yeast GCN4 protein (*58*). Our eventual goal was to perform selections on GST-fusion proteins, and so we expressed GST-GCN4 in BL21 *E. coli* cells as a positive-control bait. As a negative-control scFv, a random clone ("rand-scFv") was picked from a naive human repertoire (*59*) and expressed in the same ribosome display vector backbone. A negative-control peptide, "GST-pep" was used as nonspecific bait. Protocol optimization experiments were undertaken to maximize the amount of both enrichment and recovery of the Pluck-scFv that be could be attained. For most experiments, Pluck-scFv was diluted 1,000-fold into a background of rand-scFv, and GST-GCN4 was diluted 1,000-fold into a background of GST-pep. Our selection protocol typically achieved enrichments of several hundred-fold, and recovery of ~0.2%. This relatively low rate

of recovery is consistent with known inefficiencies inherent to the ribosome display technology (*60*).

We incorporated a system of quality-control measures to en- sure the success of each round of hidden Markov model (HMM) scFv library selection. First, we spiked Pluck-scFv into our HMM scFv library and GST-GCN4 into our selection bait (GST- PVRL4), both at a dilution of 1:1,000. In this way, the efficiency of enrichment and recovery for each selection could be quantitatively monitored using a probe specific for the Pluck-scFv control. If enrichment or recovery of Pluck-scFv was below a threshold, then the selection was considered a failure and repeated. For our selections, we required enrichment of Pluck-scFv to be at least 50- fold and the recovery of Pluck-scFv be at least 0.04%. Second, degradation of mRNA transcripts is a concern with ribosome display, and so we used TaqMan probes targeting the constant 3′ and 5′ ends of the scFv transcript. In the absence of mRNA degradation, these two signals arise with equal strength. The distal, 5′ signal is differentially diminished by degradation, and so the ratio of the two signals can be used to measure the amount of degradation that occurred during the selection. If the ratio of the 5′ signal to the 3′ signal was below our threshold of 1:5, the selection was considered a failure and repeated.

Illumina Sequencing

Libraries for Illumina sequencing were pre- pared by two rounds of PCR amplification to add the Illumina adapters and barcode sequences. Libraries were PCR-amplified from the in vitro transcription template DNA using the TaKaRa EX HS kit (Clontech). The conditions for the first round of PCR were: 1× TaKaRa EX HS Buffer, 0.2 mM dNTP, 0.4 μM IS7_L3F_PE primer, 0.4 μM IS8_H3R_PE_Multi primer, 0.5 μL TaKaRa Ex HS enzyme, and 1 μL of template per 50-μL reaction. The thermal profile was: [(i) 98°C for 10 s, (ii) 50°C for 30 s, (iii) 72°C for 1 min 30 s]

× 10 cycles, (iv) 72°C for 7 min. The conditions for the second round of PCR were: 1× TaKaRa EX HS Buffer, 0.2 mM dNTP, 0.5 μM of IS4_L3F_PE primer, 0.5 μM of the barcoding primer, 0.5 μL TaKaRa Ex HS enzyme, and 1 μL of the first round PCR product per 50-μL reaction. The thermal profile was: [(i) 98 °C for 10 s, (ii) 60 °C for 30 s, (iii) 72 °C for 1 min 30 s] × 10 cycles, (iv) 72 °C for 7 min.

As the complexity of the libraries is expected to decrease significantly with each round of selection, we divided the contribution of each library by two for each round of enrichment undergone. For example, if we added 100 ng of input library product to the multiplex pool, then we would add 50 ng of round 1 selected library, 25 ng of round 2 selected library, 12.5 ng of round 3 selected library, and so on. All second-round PCR products were gel-purified before sequencing on an Illumina HiSeq 2000 instrument.

Analysis of High-Throughput Sequencing Results

All reads were separated into samples according to the barcode sequence by the standard Illumina software. Framework sequences were trimmed according to the following rules: L3 and H2 reads were truncated to their respective lengths (36 nt and 39 nt, respectively). H3 reads were trimmed of 5′ and 3′ framework sequences with an error rate of 0.2 using cutadapt (*61*). Reads were then aligned to con- sensus sequences with up to two mismatches using bowtie software (*62*): First, all of the reads in a sample were tallied and an index was built for each sample. Second, each read in a sample was aligned globally against that sample's index with up to two mismatches allowed. The alignment with the highest tally (i.e., the read that occurred most frequently in that sample) was chosen as the consensus sequence for that read. Finally, reads that contain wildcards ("N") or stop codons were discarded. The paired L3-H3 or H2-H3 reads were then joined and the fre- quency of unique pairs was tallied. For paired L3-H3, we obtained 2.58 ×

10^7, 2.12×10^7, 9.82×10^6, 1.08×10^6, 5.19×10^5, and 5.33×10^5 total reads for the input library, round 1, round 2, round 3, round 4, and GCN4 selections, respectively. After applying our filtering algorithm, we obtained 1.60×10^7, 1.97×10^7, 8.66×10^6, 8.77×10^5, 4.71×10^5, and 4.97×10^5 reads, respectively. For paired H2-H3, we obtained 2.02×10^7, 1.68×10^7, 7.55×10^6, 8.08×10^5, 3.90×10^5, and 4.10×10^5 total reads for the input library, round 1, round 2, round 3, round 4, and GCN4 selections, respectively. After applying our filtering algorithm, we obtained 1.89×10^7, 7.16×10^6, 6.71×10^6, 6.72×10^5, 3.55×10^5, and 2.94×10^5 reads, respectively. After four rounds of selection the median read depth of the top 10 L3-H3 paired clones was 55.5 and the median read depth of the top 10 H2-H3 paired clones was 289.5. In each of the libraries, there is a long tail of clones that are sequenced only once.

Live-Cell FACS Analysis

Telomerase-large T-antigen–immortalized human mammary epithelial cells (TL-HMECs) were transduced with retroviral constructs expressing human PVRL4 or control (empty vector). For labeling with in vitro-translated scFvs, cells were dissociated from the tissue-culture plate with enzyme-free cell dissociation buffer (Invitrogen), resuspended in Stain buffer (BD Biosciences), and filtered through a 35-μm nylon mesh cell strainer (BD Biosciences). Cells were incubated with in vitro- translated FLAG-tagged scFvs at a 1:100 dilution or anti-PVRL4 mouse monoclonal antibody (R&D Systems) for 30 min on ice, washed twice with Stain buffer, and incubated with M2 anti- FLAG antibody (Sigma) at a 1:100 dilution for 30 min on ice. Labeled cells were washed twice and incubated with Alexa Fluor 488-conjugated goat–anti-mouse secondary antibody (Invitrogen) at 1:500 dilution for 30 min on ice. After a final series of washes, cells were resuspended in Stain buffer. Fluorescent signal was measured on LSR II FACS Analyzer (BD Bio- sciences) and analyzed with FlowJo software.

Chapter 3:
Development of a synthetic human virome for comprehensive serological profiling

George J. Xu[1-4,#], Tomasz Kula[3-5,#], Qikai Xu[3,4], Mamie Z. Li[3,4], Suzanne D. Vernon[6], Thumbi Ndung'u[7,8,9,10], Kiat Ruxrungtham[11], Jorge Sanchez[12], Christian Brander[13], Raymond T. Chung[14], Kevin C. O'Connor[15], Bruce Walker[8,9], H. Benjamin Larman[16], Stephen J. Elledge[3,4,]

1. Program in Biophysics, Harvard University, Cambridge, MA.
2. Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA.
3. Division of Genetics, Department of Medicine, Howard Hughes Medical Institute, Brigham and Women's Hospital, Boston, MA.
4. Department of Genetics, Harvard University Medical School, Boston, MA.
5. Program in Biological and Biomedical Sciences, Harvard University, Cambridge, MA.
6. Solve ME/CFS Initiative, Los Angeles, CA.
7. KwaZulu-Natal Research Institute for Tuberculosis and HIV, Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa.
8. HIV Pathogenesis Programme, Doris Duke Medical Research Institute, Nelson R. Mandela School of Medicine, Durban, South Africa.
9. The Ragon Institute of Massachusetts General Hospital, Massachusetts Institute of Technology, and Harvard University, Cambridge, MA.

10. Max Planck Institute for Infection Biology, Chariteplatz, D-10117 Berlin, Germany.
11. Vaccine and Cellular Immunology Laboratory, Department of Medicine, Faculty of Medicine; and Chula-Vaccine Research Center, Chulalongkorn University, Bangkok, Thailand.
12. Asociación Civil IMPACTA Salud y Educación, Lima, Peru.
13. IDS Research Institute-IrsiCaixa and AIDS Unit, Hospital Germans Trias i Pujol, Universitat Autònoma de Barcelona, Badalona, Spain Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.
14. Division of Gastroenterology, Massachusetts General Hospital, Boston, MA.
15. Dept. of Neurology, Yale School of Medicine, New Haven, CT.
16. Department of Pathology, Johns Hopkins University, Baltimore, MD.

# These authors contributed equally to this work.

Acknowledgements

We thank Elizabeth Unger and Supranee Buranapraditkun for providing reagents and Kai Wucherpfennig (Harvard) and Hidde Ploegh (MIT) for critical reading of the manuscript, and TWIST Biosciences for providing access to their advanced oligonucleotide synthesis technology. The cohort in Durban, South Africa was funded by the NIH (R37AI067073) and the International AIDS Vaccine Initiative (UKZNRSA1001). T.N. received additional funding from the South African Research Chairs Initiative, the Victor Daitz Foundation and an International Early Career Scientist Award from the Howard Hughes Medical Institute. RTC was funded by grants NIH DA033541 and AI082630. C.B and J.S. were supported by NIH N01-AI-30024 and

N01-Al-15422, NIH-NIDCR R01 DE018925-04, the HIVACAT program and CUTHIVAC 241904. K.R. is supported by TRF Senior Research Scholar, the Thailand Research Fund; and the Chulalongkorn University Research Professor Program, Thailand and NIH grant N01-A1-30024. G.J.X. and T.K. were supported by the NSF Graduate Research Fellowships Program. S.J.E. and B.W. are Investigators with the Howard Hughes Medical Institute. G.J.X., T.K., H.B.L., and S.J.E. are inventors on a patent application (PCT Application No. PCT/US14/70902) filed by The Brigham and Women's Hospital, Inc. that covers the use of phage display libraries to detect antiviral antibodies.

Author Contribution

S.J.E. conceived and supervised the project. H.B.L. and Q.K. designed the initial virus library. M.Z.L. constructed the libraries. The new library was designed by G.J.X. and constructed by M.Z.L. G.J.X. designed and constructed the mutagenesis library. G.J.X. and T.K. performed the experiments. G.J.X. developed the computational pipeline to process the raw sequencing data. G.J.X. and T.K. performed the data analysis. The manuscript was prepared by G.J.X. and edited by T.K., H.B.L. and S.J.E. All other authors provided samples and feedback on the manuscript.

Adapted from G. Jing Xu, *et al*. "Construction of a Rationally Designed Antibody Platform for Sequencing-assisted Selection." *Science*. *In press*.

Abstract

The human virome plays important roles in health and immunity. However, current methods for detecting viral infections and antiviral responses have limited throughput and coverage. Here, we present VirScan, a high-throughput method to comprehensively analyze antiviral antibodies using immunoprecipitation and massively parallel DNA sequencing of a bacteriophage library displaying proteome-wide peptides from all human viruses. We assayed over 10^8 antibody-peptide interactions in 569 humans across four continents, nearly doubling the number of previously established viral epitopes. We detected antibodies to an average of 10 viral species per person and 84 species in at least two individuals. Although rates of specific virus exposure were heterogeneous across populations, antibody responses targeted strikingly conserved “public epitopes” for each virus, suggesting that they may elicit highly similar antibodies. VirScan is a powerful approach for studying interactions between the virome and the immune system.

Introduction

The collection of viruses found to infect humans (the "human virome”) can have profound effects on human health (*63*). In addition to directly causing acute or chronic illness, viral infection can also alter host immunity in more subtle ways, leaving an indelible footprint on the immune system (*64*). For example, latent herpesvirus infection has been shown to confer symbiotic protection against bacterial infection in mice through prolonged production of interferon-γ and systemic activation of macrophages (*65*). This interplay between virome and host immunity has also been implicated in the pathogenesis of complex diseases such as type 1 diabetes, inflammatory bowel disease, and asthma (*66*). Despite this growing appreciation for the

importance of interactions between the virome and host, a comprehensive method to systematically characterize these interactions has yet to be developed (*67*).

Viral infections can be detected by serological- or nucleic acid-based methods (*68*). However, nucleic acid tests fail in cases where viruses have already been cleared after causing or initiating tissue damage and can miss viruses of low abundance or viruses not normally present in the sampled fluid or surface. In contrast, humoral responses to infection typically arise within two weeks of initial exposure and can persist over years or decades (*21*). Tests detecting antiviral antibodies in peripheral blood can therefore identify ongoing and cleared infections. However, current serological methods are predominantly limited to testing one virus at a time and are therefore only employed to address specific clinical hypotheses. Scaling serological analyses to encompass the complete human virome poses significant technical challenges, but would be of great value for better understanding host-virus interactions, and would overcome many of the limitations associated with current clinical technologies. In this work, we present VirScan, a programmable, high-throughput method to comprehensively analyze antiviral antibodies using immunoprecipitation and massively parallel DNA sequencing of a bacteriophage library displaying proteome-wide coverage of peptides from all human viruses.

Results

The VirScan platform

VirScan utilizes the Phage Immunoprecipitation sequencing (PhIP-seq) technology previously developed in our laboratory (*19*). Briefly, we used a programmable DNA microarray to synthesize 93,904 200-mer oligonucleotides, encoding 56-residue peptide tiles, with 28 residue overlaps, that together span the reference protein sequences (collapsed to 90% identity)

of all viruses annotated to have human tropism in the UniProt database (Figure 8a and b) (*69*). This library includes peptides from 206 species of virus and over 1,000 different strains. We cloned the library into a T7 bacteriophage display vector for screening (Figure 8c).

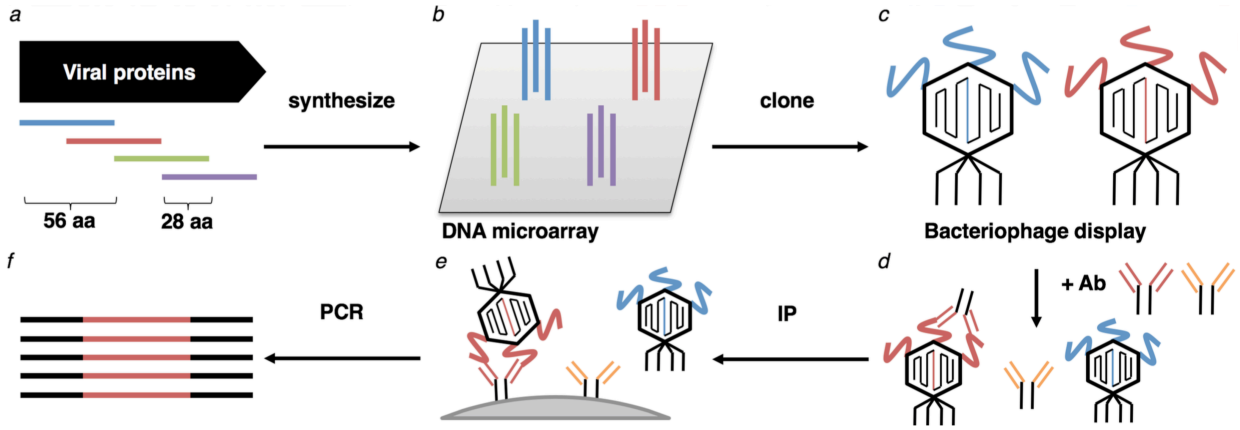

Figure 8. Construction of the virome peptide library and VirScan screening procedure. (*a*) The virome peptide library consists of 93,904 56 amino acid peptides tiling, with 28 amino acid overlap, across the proteomes of all known human viruses. (*b*) 200 nt DNA sequences encoding the peptides were printed on a releasable DNA microarray. (*c*) The released DNA was amplified and cloned into a T7 phage display vector and packaged into virus particles displaying the encoded peptide on its surface. (*d*) The library is mixed with a sample containing antibodies that bind to their cognate peptide antigen on the phage surface. (*e*) The antibodies are immobilized and unbound phage are washed away. (*f*) Finally, amplification of the bound DNA and high throughput sequencing of the insert DNA from bound phage reveals peptides targeted by sample antibodies. Abbreviations: aa, amino acid; Ab, antibody; IP: immunoprecipitation.

To perform a screen, we incubate the library with a serum sample containing antibodies, recover the antibodies using a mixture of protein A and G coated magnetic beads, and remove unbound phage particles by washing (Figure 8d and e). Finally, we perform PCR and massively parallel sequencing on the phage DNA to quantify enrichment of each library member due to antibody binding (Figure 8f). Each sample is screened in duplicate to ensure reproducibility.

VirScan requires only 2 μg of immunoglobulin (<1 μL of serum) per sample and can be automated on a 96-well liquid handling robot (*70*). PCR product from 96 immunoprecipitations can be individually barcoded and pooled for sequencing, reducing the cost for a comprehensive viral antibody screen to approximately $25 per sample.

Following sequencing, we tally the read count for each peptide before ("input") and after ("output") immunoprecipitation. We then fit a zero-inflated generalized Poisson model to the distribution of output read counts for each input read count and regress the parameters as a function of input read count (Figure 9).

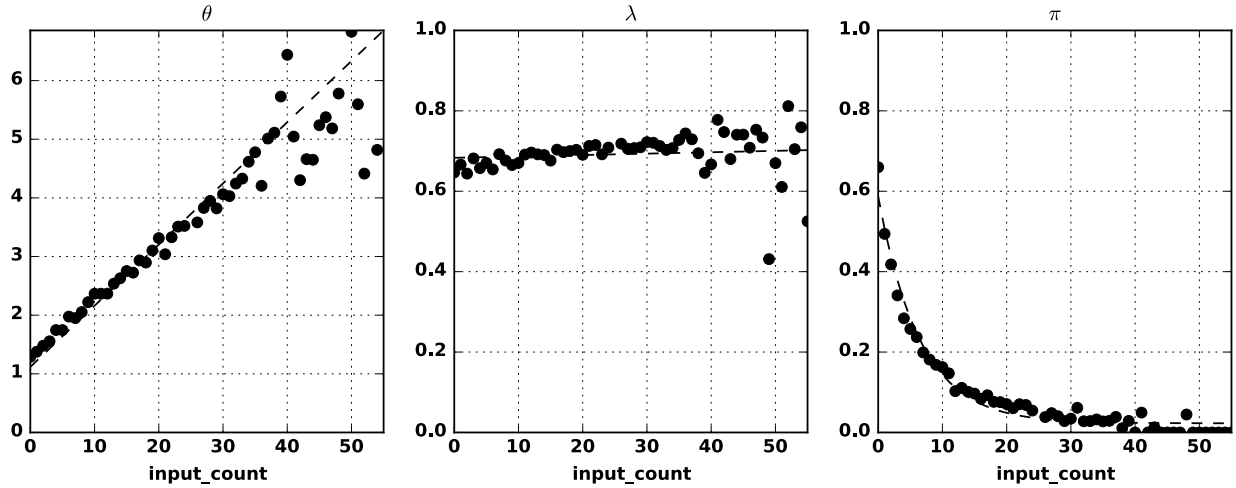

Figure 9. Zero inflated generalized poisson (ZIGP) parameters regressed on input count. Each scatter plot depicts the maximum likelihood estimates for the ZIGP parameters as a function of the input count (horizontal axis; see Materials and Methods). Dashed lines are least-squares linear regressions for θ and λ, and least-squares exponential regression for π.

Using this model, we calculate a $-\log_{10}$(p-value) for the significance of each peptide's enrichment. Finally, we call a peptide significantly enriched if its $-\log_{10}$(p-value) is greater than the reproducibility threshold of 2.3 in both replicates (Figure 10).

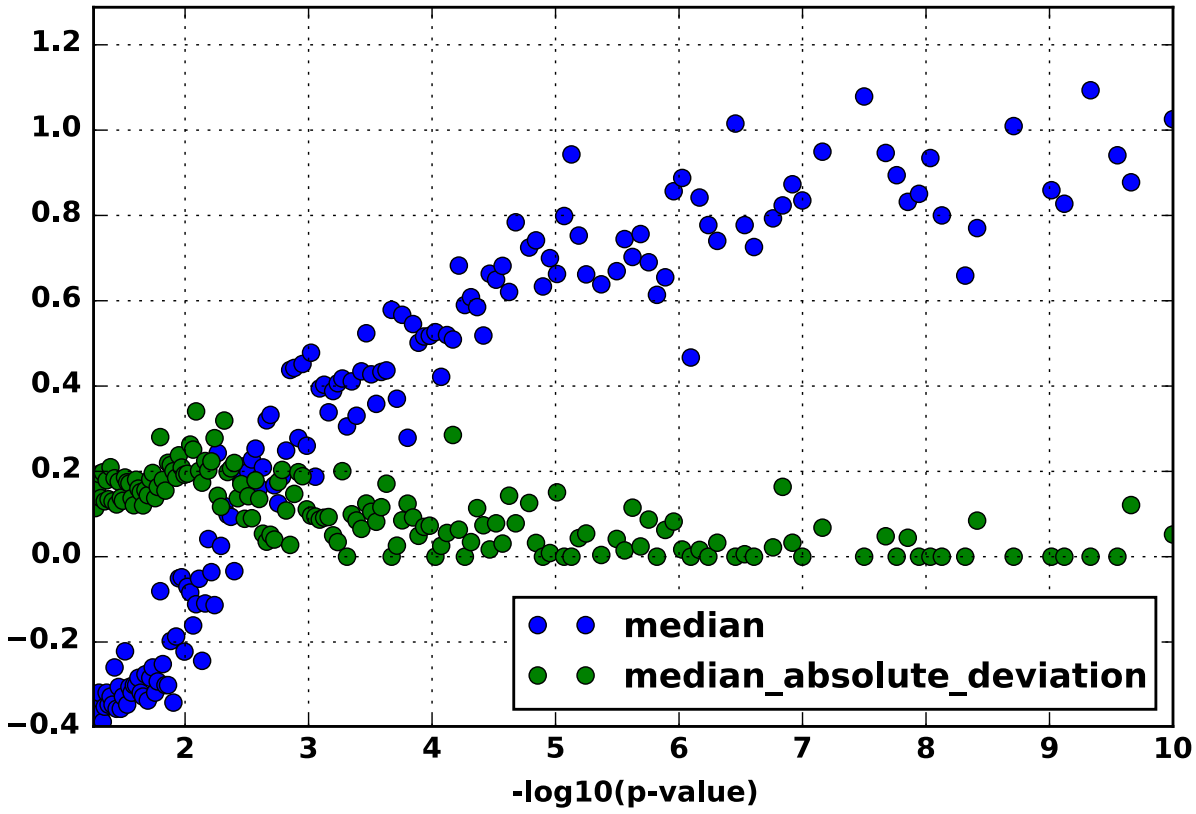

Figure 10. Reproducibility threshold. Scatterplot for median and median absolute deviation of replicate 2 -log10(p-values) whose replicate 1 -log10(p-value) falls within the window whose left edge is shown on the horizontal axis (see Materials and Methods).

Characterizing VirScan's sensitivity and specificity

Figure 11 shows the antibody profiles of a set of human viruses in sera from a typical group of individuals in a heat map format that illustrates the number of enriched peptides from each virus. We frequently detected antibodies to multiple peptides from common human viruses, such as Epstein-Barr virus (EBV), Cytomegalovirus (CMV), and rhinovirus. As expected, we observed more peptides to be enriched from viruses with larger proteomes, such as EBV and CMV, likely because there are more epitopes available for recognition. We noticed fewer enriched peptides in samples from individuals less than ten years of age compared to their

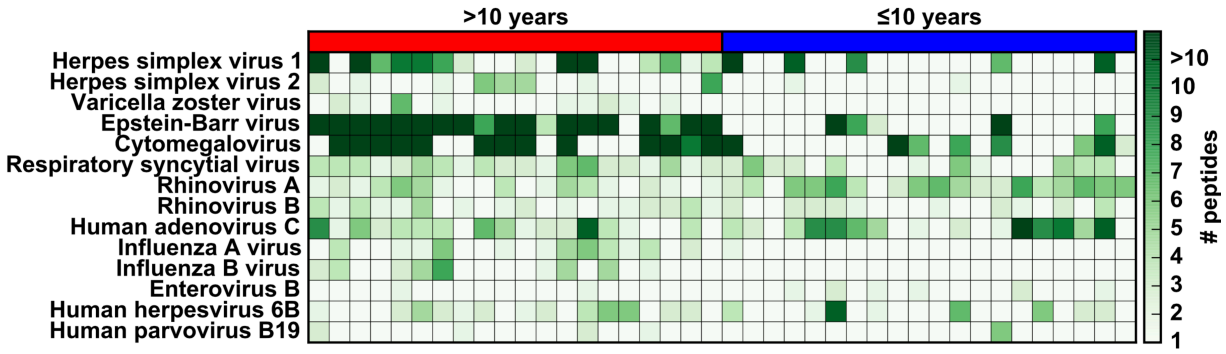

Figure 11. Antibody profile of randomly chosen group of donors to show typical assay results. Each row is a virus, each column is a sample. The label above each chart indicates whether the donors are over 10 years of age or at most 10 years of age. The color intensity of each cell indicates the number of peptides from the virus that were significantly enriched by antibodies in the sample.

geographically matched controls, in line with an accumulation of viral infections throughout adolescence and adulthood. However, there were occasional samples from young donors with very strong responses to viruses that cause childhood illness, such as Parvovirus B19 and Herpesvirus 6B, which cause the "fifth disease" and "sixth disease" of the classical infectious childhood rashes, respectively (*71*). These observations are examined in greater detail in Figure 15.

We developed a computational method to identify the set of viruses to which an individual has been exposed, based on the number of enriched peptides identified per virus. Briefly, we set a threshold number of significant non-overlapping enriched peptides for each virus. We empirically determined that a threshold of three non-overlapping enriched peptides gave the best performance for detecting Herpes simplex virus 1 compared to a commercial serologic test, described below (Table 2). For other viruses, we adjusted the threshold to account for the size of the viral proteome (Figure 12). Next, we tally the number of enriched peptides from each virus. Antibodies generated against a specific virus can cross-react with similar

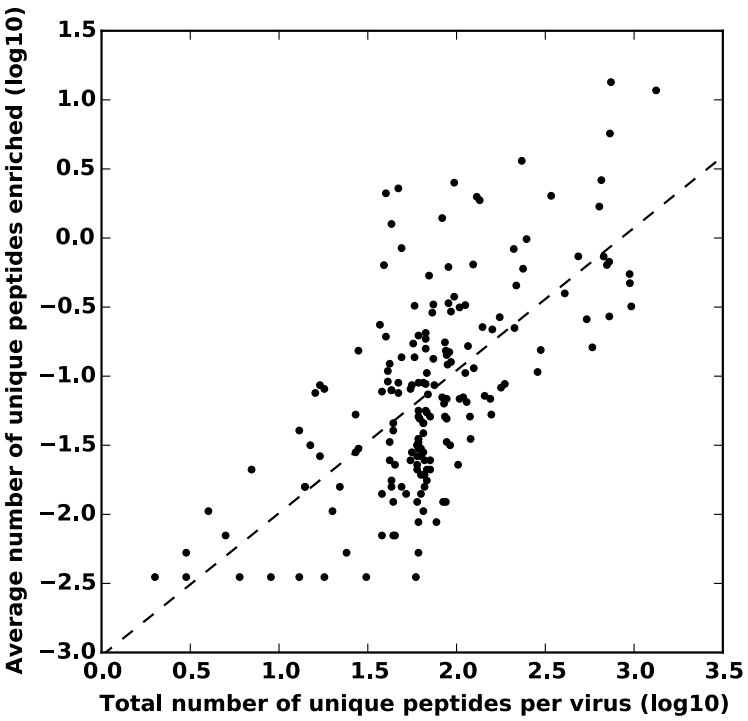

Figure 12. Correlation between virus size and number of enriched peptides. Each dot on this log-log scatterplot is a virus. The horizontal axis corresponds to the size of the virus in number of peptides. The vertical axis corresponds to the average number of peptides enriched from the virus across all samples tested. The dashed line is a least-squares best-fit curve for the data.

peptides from a related virus. This would lead to false positives because an antibody targeted to an epitope from one virus to which a donor was exposed would also enrich a homologous peptide from a related virus to which the donor may not have been exposed. In order to address this issue, we adopted a maximum parsimony approach to infer the fewest number of virus exposures that could elicit the observed spectrum of antiviral peptide antibodies. For groups of enriched peptides that share a 7 amino acid subsequence and may be recognized by a single specific antibody, we only count it as one epitope for the virus that has the greatest number of other enriched peptides. If this adjusted peptide count is greater than the threshold for that virus,

the sample is considered positive for the virus. For this analysis, we also filtered out peptides that were enriched in only one of the 569 samples to avoid spurious hits.

Using this analytical framework, we measured the performance of VirScan using serum samples from patients known to be infected or not infected with human immunodeficiency virus (HIV) and Hepatitis C virus (HCV), based on commercial ELISA and Western blot assays. For both viruses, VirScan achieves very high sensitivities and specificities of ~95% or higher (Table 2) over a wide range of viral loads (Figure 13). The viral genotype was also known for the HCV positive samples. Despite the over 70% amino acid sequence conservation among HCV genotypes (*72*), which poses a problem for all antibody-based detection methods, VirScan correctly reported the HCV genotype in 69% of the samples. We also compared VirScan to a commercially available serology test that is type specific for the highly related Herpes simplex viruses 1 and 2 (HSV1 and HSV2) (Table 1). These results demonstrate that VirScan performs well in distinguishing between closely related viruses and viruses that range in size from small (HIV and HCV) to very large (HSV1 and HSV2) with high sensitivity and specificity.

Virus	Sensitivity (*n*)	Specificity (*n*)
Hepatitis C virus	92% (26) *	97% ** (34)
Human immunodeficiency virus 1	95% (61) *	100% (33)
Herpes simplex virus 1	97% (38)	100% (6)
Herpes simplex virus 2	90% (20)	100% (24)

Table 2. Virscan's sensitivity and specificity on samples with known viral infections. Sensitivity is the percentage of samples positive for the virus as determined by VirScan out of all *n* known positives. Specificity is the percentage of samples negative for the virus by VirScan out of all *n* known negatives.

* We found that although the false negative samples did not meet our stringent cut-off for enriching multiple unique peptides, they had detectable antibodies to a recurrent epitope. By modifying the criterion to allow for samples that enrich multiple homologous peptides that share a recurrent epitope as described in the text, the sensitivity of

detecting Hepatitis C virus increases to 100% and the sensitivity for detecting HIV increases to 95%. This modified criterion does not significantly affect specificity (fig. S13).

** The one false positive was from an individual whose HCV-negative status was self-reported, but had antibodies to as many HCV peptides as 23% of the true HCV positive individuals and is likely to be HCV positive now or in the past. It is possible that this individual was exposed to HCV but cleared the infection. If true, the observed specificity for HCV is 100%.

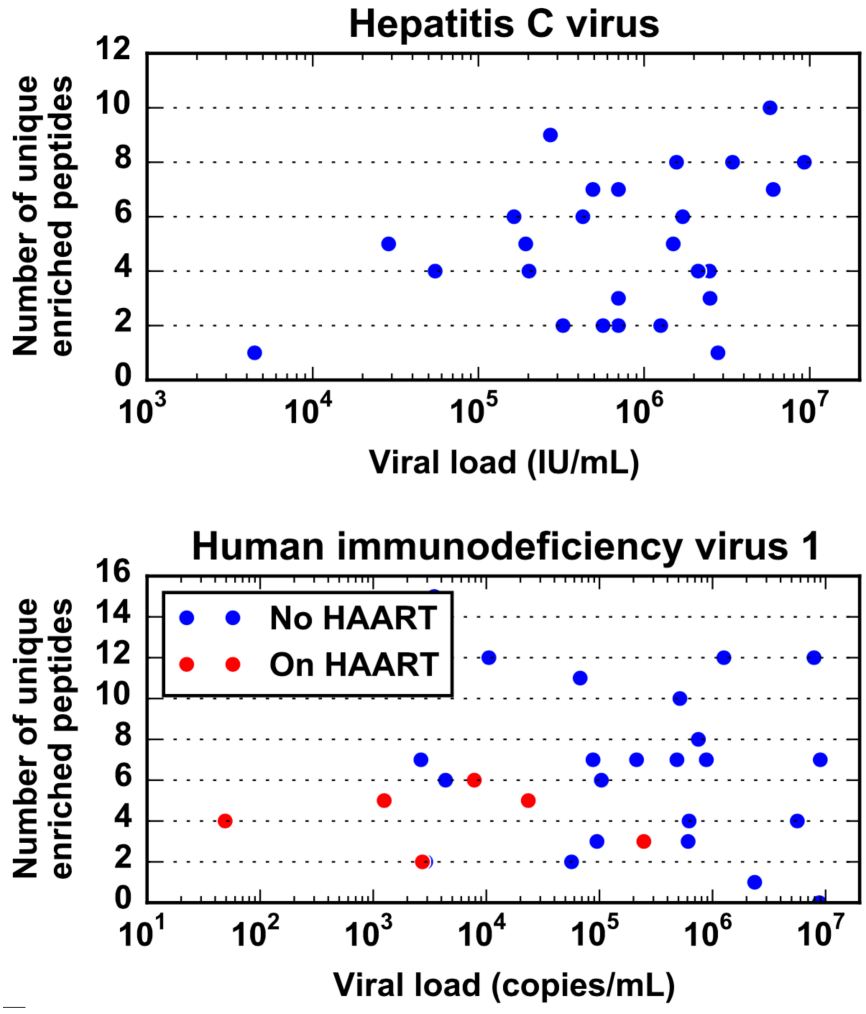

Figure 13. Scatter plot of the number of unique enriched peptides (after applying maximum parsimony filtering) detected in each sample against the viral load in that sample. Data are shown for the HCV positive and HIV positive samples for which we were able to obtain viral load data. For the HIV positive samples, red dots indicate samples from donors currently on highly active anti-retroviral therapy at the time the sample was taken, whereas blue dots indicate different donors prior to undergoing therapy.

Population-level analysis of viral exposures

After ascertaining the performance of VirScan for a panel of viruses, we undertook a large-scale screening of samples with unknown exposure history. Using our multiplex approach, we assayed over 106 million antibody-peptide interactions using samples from 569 human donors in duplicate. We detected antibody responses to an average of 10 species of virus per sample (Figure 14).

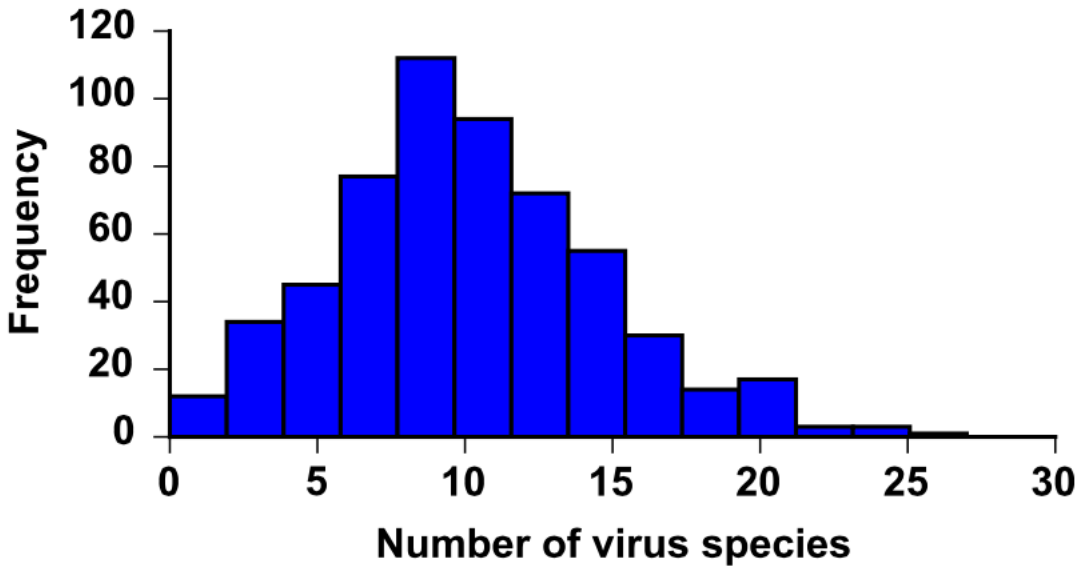

Figure 14. Distribution of number of viruses detected in each sample. The histogram depicts the frequency of samples binned by the number of virus species detected by VirScan. The mean and median of the distribution are both approximately 10 virus species.

Each person is likely exposed to multiple distinct strains of some viral species. We detected antibody responses to 62 of the 206 species of virus in our library in at least 5 individuals, and 84 species in at least 2 individuals. The most frequently detected viruses are generally those known to commonly infect humans (Table 3). We occasionally detected what appear to be false positives that may be due to antibodies that cross react with non-viral peptides. For example, 29% of the samples positive for Cowpox virus were right at the threshold

	Detection Frequency	Above Minimum Threshold	Most Recurrent Peptide	Fraction Peptides Recurrent	Number Unique Peptides Recurrent
Human herpesvirus 4	87.1%	98.5%	87.4%	1.4%	13
Rhinovirus B	71.8%	52.7%	96.4%	5.0%	5
Human adenovirus C	71.8%	80.2%	71.6%	0.8%	4
Rhinovirus A	67.3%	59.1%	99.0%	4.6%	8
Human respiratory syncytial virus	65.7%	67.0%	86.2%	5.7%	4
Human herpesvirus 1	54.4%	87.5%	89.9%	1.1%	6
Influenza A virus	53.4%	57.0%	55.2%	0.1%	1
Human herpesvirus 6B	52.8%	66.3%	61.3%	0.7%	4
Human herpesvirus 5	48.5%	96.7%	95.3%	0.9%	19
Influenza B virus	40.5%	55.2%	51.2%	1.7%	4
Poliovirus	33.7%	40.4%	81.7%	2.0%	2
Human herpesvirus 3	24.3%	54.7%	77.3%	1.0%	4
Human adenovirus F	20.4%	17.5%	81.0%	0.4%	3
Human adenovirus B	16.8%	38.5%	71.2%	0.6%	3
Human herpesvirus 2	15.5%	75.0%	85.4%	0.7%	6
Enterovirus A	15.2%	12.8%	70.2%	2.3%	3
Enterovirus B	13.3%	7.3%	95.1%	3.3%	5
Mamastrovirus 1	9.4%	24.1%	55.2%	0.7%	2
Human herpesvirus 7	9.1%	42.9%	92.9%	0.4%	4
Norwalk virus	8.7%	25.9%	96.3%	1.2%	3
Human adenovirus D	8.4%	38.5%	50.0%	0.4%	3
Human parainfluenza virus 3	7.4%	21.7%	47.8%	1.6%	2
Cowpox virus	7.1%	9.1%	36.4%	0.1%	1
Human adenovirus A	6.5%	35.0%	55.0%	0.5%	2
Human metapneumovirus	5.2%	43.8%	43.8%	2.8%	4
Human coronavirus HKU1	4.5%	0.0%	42.9%	0.2%	3
Human herpesvirus 6A	4.2%	30.8%	46.2%	0.4%	4
Alphapapillomavirus 9	4.2%	30.8%	61.5%	0.5%	3
Human parvovirus B19	3.9%	25.0%	75.0%	1.5%	3
Aichivirus A	3.9%	33.3%	66.7%	2.6%	5
Hepatitis B virus	3.6%	9.1%	18.2%	0.0%	0
Betapapillomavirus 1	3.2%	0.0%	40.0%	0.1%	1
Influenza C virus	2.9%	33.3%	55.6%	0.2%	2
Human coronavirus NL63	2.9%	0.0%	55.6%	1.1%	3
Human herpesvirus 8	2.6%	50.0%	50.0%	0.5%	4
Rubella virus	2.6%	12.5%	50.0%	1.5%	2
Human adenovirus E	2.3%	14.3%	71.4%	0.5%	1
Hepatitis E virus	1.9%	0.0%	33.3%	0.4%	3
Torque teno virus	1.6%	0.0%	60.0%	0.9%	3
Hepatitis C virus	1.6%	80.0%	13.1%	0.0%	0
Measles virus	1.6%	20.0%	80.0%	2.2%	3
Alphapapillomavirus 10	1.6%	0.0%	80.0%	1.1%	3
Human parainfluenza virus 4	1.6%	0.0%	80.0%	6.3%	3
Eastern equine encephalitis virus	1.3%	0.0%	75.0%	0.7%	1
Rotavirus A	1.3%	0.0%	50.0%	0.1%	1

Table 3. Detection frequency of all viruses detected in at least 4 (>1%) of the 303 donors residing in the United States. Known HIV-positive and HCV-positive samples were excluded from this analysis. The "Detection Frequency" column shows the percentage of the 303 US samples that were positive for each virus. Of the samples that are positive for each virus, the "Above Minimum Threshold" column shows the percentage that enriched more unique peptides than just the minimum threshold for that virus (Fig S3), and the "Most Recurrent Peptide" column shows the percentage that enriched the most recurrent peptide for that virus. The "Number Unique Peptides Recurrent" column shows the number of unique peptides (peptides that do contain the identical subsequences of 7 amino acids or longer) from that virus that are enriched in at least 30% of the samples that are positive for that virus. The "Fraction Peptides Recurrent" column shows the total number of recurrent peptides from a virus divided by the number of all peptides from that virus.

of detection and had antibodies against a peptide from the C4L gene that shares an eight amino acid sequence ('SESDSDSD') with the Clumping Factor B protein from *Staphylococcus aureus*, against which humans are known to generate antibodies (*73*). This will become less of an issue when we test more examples of sera from patients with known infections to determine the set of likely antigenic peptides for a given virus. However, the fact that we do not detect high rates of very rare or virulent viruses strengthens our confidence in VirScan's specificity.

We frequently detected antibodies to rhinovirus and respiratory syncytial virus, which are normally found only in the respiratory tract, indicating that VirScan using blood samples is still able to detect viruses that do not cause viremia. We also detected antibodies to influenza, which is normally cleared, and poliovirus, to which most people in modern times generate antibodies through vaccination. Since the original antigen is no longer present, we are likely detecting antibodies secreted by long-lived memory B cells (*74*).

We detected antibodies to certain viruses less frequently than expected based on previous seroprevalence studies using optimized serum ELISA assays. For example, the frequency at

which we detect influenza (53.4%) and poliovirus (33.7%) is lower than expected given that the majority of the population has been exposed to or vaccinated against these viruses. This may be due to reduced sensitivity because of a gradual narrowing and decrease of the long-lived B cell response in the absence of persistent antigen. We also rarely detected antibody responses to small viruses such as JC virus and Torque Teno virus, which are frequently detected using specific tests. We believe that the disparity is due to low titers of antibodies to unmodified, linear epitopes from these viruses. For example, serum antibodies against the major capsid protein of JCV are reported to only recognize conformational epitopes (*75*). Finally, the frequency of detecting varicella zoster virus (chicken pox) antibodies is also lower than expected (24.3%), even though the frequency of detecting other latent herpesviruses, such as Epstein-Barr virus (87.1%) and cytomegalovirus (48.5%), is similar to the prevalence reported in epidemiological studies (*76*–*78*). This may reflect differences in how frequently these viruses shed antigens that stimulate B cell responses or a more limited humoral response that relies on epitopes that cannot be detected in a 56-residue peptide. It might also be possible to increase the sensitivity of detection of these viral antibodies by stimulating memory B cells in vitro to probe the history of infection more deeply.

To assess differences in viral exposure between populations, we split the samples into different groups based on age, HIV status, and geography. We first compared results from children under the age of ten to adults within the United States (HIV-positive individuals were excluded from this analysis) (Figure 15A). Fewer children were positive for most viruses, including Epstein-Barr virus, HSV1, HSV2, and influenza virus, which is consistent with our preliminary observations comparing the number of enriched peptides (Figure 11). In addition to the fact that children may generate lower antibody titers in general, these younger donors

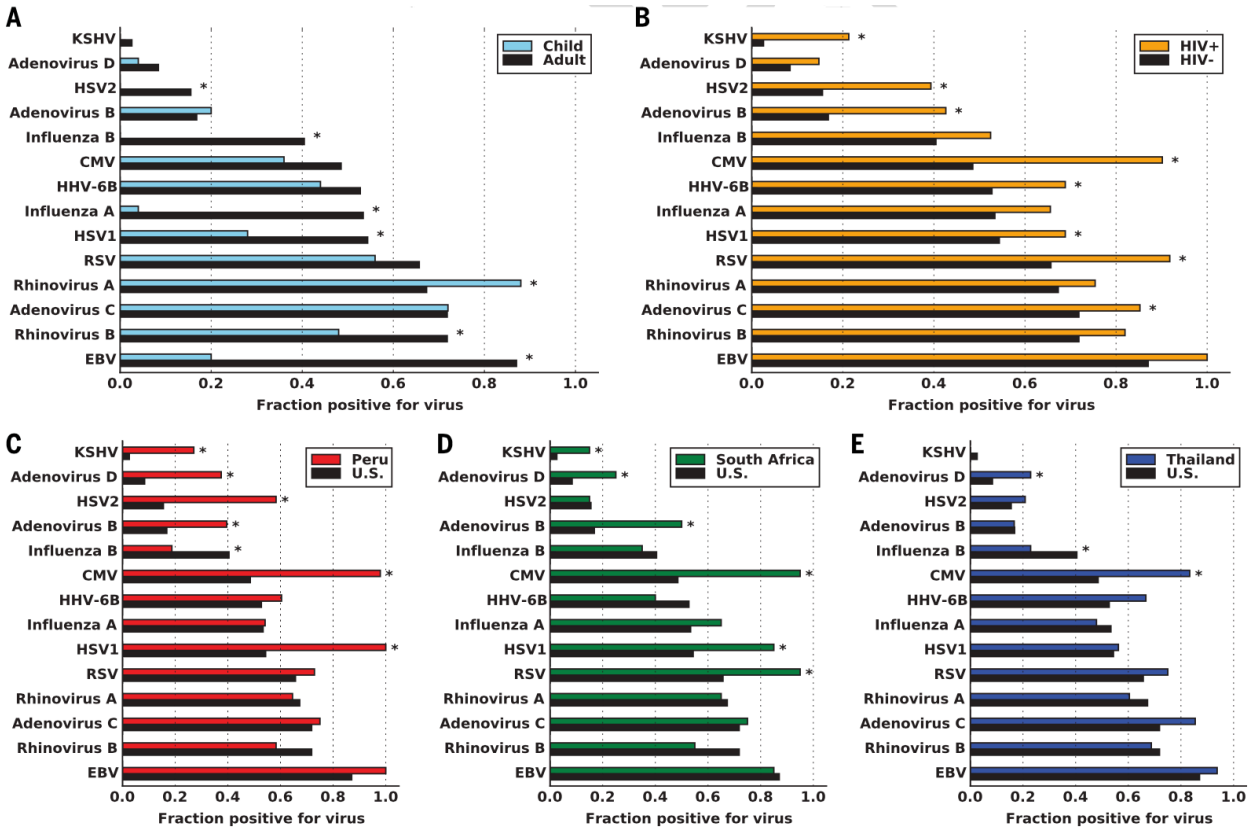

Figure 15. Population stratification of the human virome immune response. The bar graphs depict the differences in exposure to viruses between donors who are **(A)** less than ten years of age versus over ten years of age, **(B)** HIV positive versus HIV negative, **(C)** residing in Peru versus residing in the United States, **(D)** residing in South Africa versus residing in the United States, and **(E)** residing in Thailand versus residing in the United States. Asterisks indicate false discovery rate < 0.05.

probably have not yet been exposed to certain viruses, for example HSV2 which is sexually transmitted (*79*).

When comparing results from HIV positive to HIV negative samples, we found more of the HIV positive samples to also be seropositive for additional viruses, including HSV2, CMV, and Kaposi's sarcoma-associated herpesvirus (KSHV) (false discovery rate $q < 0.05$, Figure 15B). These results are consistent with prior studies indicating higher risk of these co-infections in HIV positive patients (*80–82*). Patients with HIV may engage in activities that put them at

higher risk for exposure to these viruses. Alternatively, these viruses may increase the risk of HIV infection. HIV infection may reduce the immune system's ability to control reactivation of normally dormant resident viruses or to prevent opportunistic infections from taking hold and triggering a strong adaptive immune response.

Finally, we compared the evidence of viral exposure between samples taken from adult HIV-negative donors residing in countries from four different continents (the United States, Peru, Thailand, and South Africa). In general, donors outside the United States had higher frequencies of seropositivity (Figure 15C-E). For example, cytomegalovirus antibodies were found in significantly higher frequencies in samples from Peru, Thailand, and South Africa. Other viruses, such as Kaposi's sarcoma-associated herpesvirus and HSV1 were detected more frequently in donors from Peru and South Africa, but not Thailand. The observed detection frequency of different adenovirus species varies across populations. Adenovirus C seropositivity was found at similar frequencies in all regions, but Adenovirus D seropositivity was generally higher outside the United States, while Adenovirus B seropositivity was higher in Peru and South Africa, but not in Thailand. The higher rates of virus exposure outside the United States could be due to differences in population density, cultural practices, sanitation, or genetic susceptibility. Interestingly, Influenza B seropositivity was more common in the United States compared to other countries, especially Peru and Thailand. The global incidence of Influenza B is much lower than Influenza A but the standard influenza vaccination contains both Influenza A and B strains, so the elevated frequency of individuals with seroreactivity may be due to higher rates of influenza vaccination in the United States. Other viruses, such as Rhinovirus and Epstein-Barr virus, were detected at very similar frequencies in all the geographic regions.

Analysis of viral epitope determinants

After analyzing responses on the whole virus level, we focused our attention on the specific peptides targeted by these antibodies. We detected antibodies to a total of 8,425 peptides in at least 2 samples, and 15,052 in at least 1 sample. Because of the presence of many related peptides in our library and the Immune Epitope Database (IEDB), for the following analysis we consider a peptide unique only if it does not contain a continuous 7-residue subsequence, the estimated size of a linear epitope, in common with another peptide. Analyzed as such, our VirScan database nearly doubles the 1,559 unique human B cell epitopes from human viruses in the IEDB (*83*). The epitopes identified in our unbiased analysis demonstrate a significant overlap with those contained in the IEDB ($p < 10^{-30}$, Fisher's exact text, Figure 16). The amount of overlap is even greater for epitopes from viruses that commonly cause infection. We would likely have detected even more antigenic peptides in common with the IEDB if we had tested more samples from individuals infected with rare viruses. We next analyzed the amino acid composition of recurrently enriched peptides. Enriched peptides tend to have more proline and charged amino acids and fewer hydrophobic amino acids, which is consistent with a previous analysis of B cell epitopes in the IEDB (Figure 17) (*84*). This trend likely reflects enrichment for amino acids that are surface exposed or can form stronger interactions with antibodies.

B cell responses target highly similar viral epitopes across individuals

We compared the profile of peptides recognized by the antibody response in different individuals. We found that for a given protein, each sample generally only had strong responses against one to three immunodominant peptides (Figure 18). Surprisingly, we found that the vast

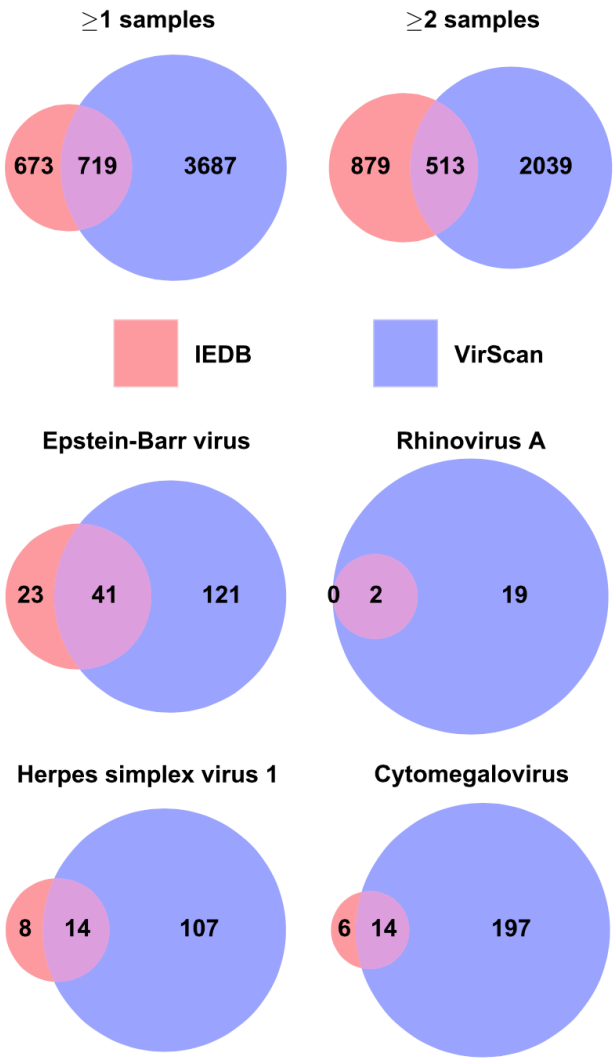

Figure 16. (Upper portion) Overlap between enriched peptides detected by VirScan and human B cell epitopes from viruses in IEDB. The entire pink circle represents the 1,392 groups of non-redundant IEDB epitopes that are also present in the VirScan library (out of 1,559 clusters total). The overlap region represents the number of groups with an epitope that is also contained in an enriched peptide detected by VirScan. The purple only region represents the number of non-redundant enriched peptides detected by VirScan that do not contain an IEDB epitope. Data are shown for peptides enriched in at least one (left) or at least two (right) samples. (Lower portion) Overlap between enriched peptides detected by VirScan and human B cell epitopes in IEDB from common human viruses. The regions represent the same values as the top except only epitopes corresponding to the indicated virus are considered, and only peptides from that virus that were enriched in at least two samples were considered.

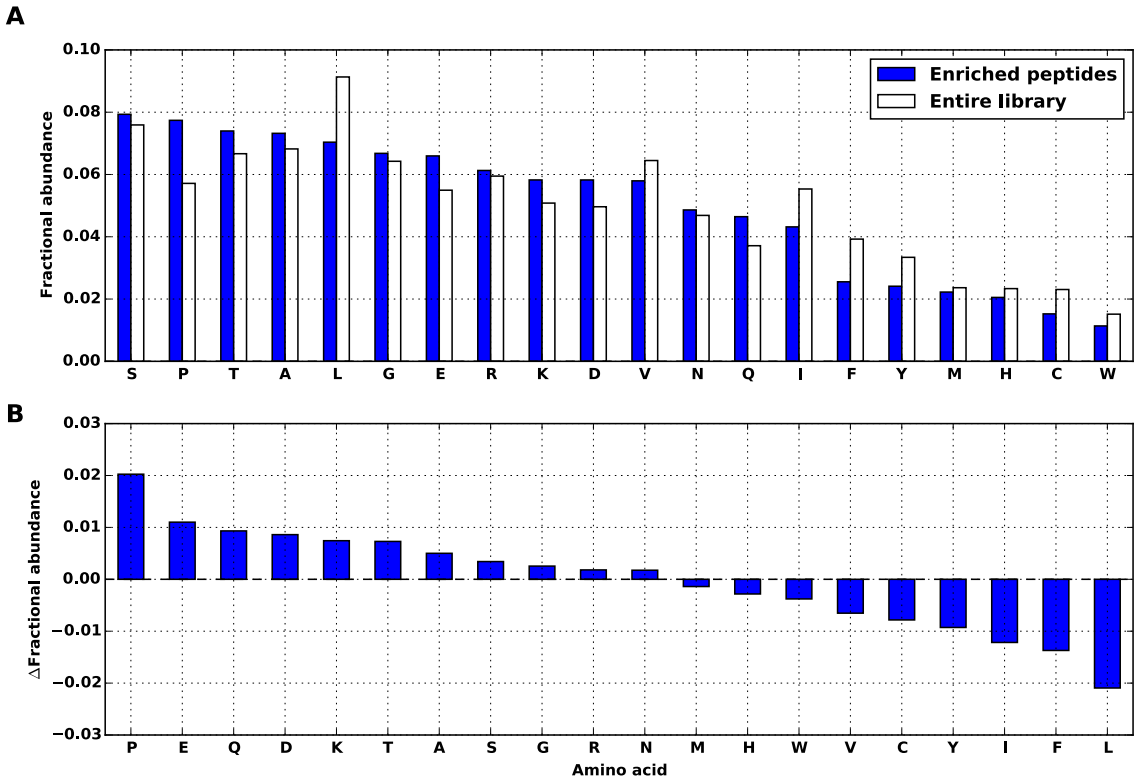

Figure 17. Amino acid composition of enriched peptides. (A) Bar graph of the fractional abundance of each amino acid in the entire virome peptide library or peptides enriched in at least 2 samples. (B) Bar graph of the fractional abundance of each amino acid in peptides enriched in at least 2 samples subtracted by the abundance in the entire library.

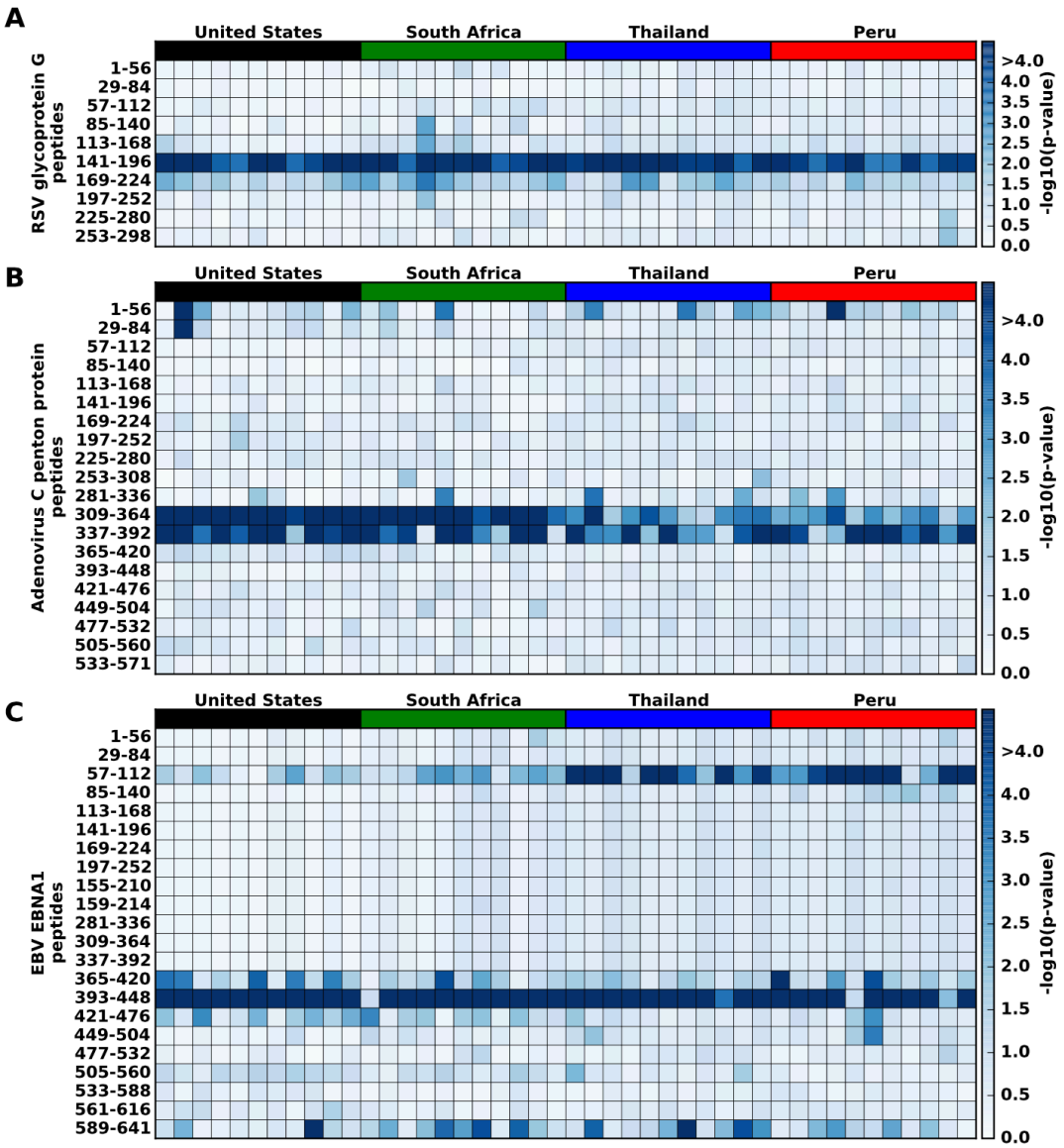

Figure 18. The human anti-virome response recognizes a similar spectrum of peptides among infected individuals. In the heatmap charts, each row is a peptide tiling across the indicated protein and each column is a sample. The colored bar above each column, labeled at the top of the figure, indicates the country of origin for that sample. The samples shown are a subset of individuals with antibodies to at least one peptide from the protein. The color intensity of each cell corresponds to the $-\log_{10}$(p-value) measure of significance of enrichment for a peptide in a sample (greater values indicates stronger antibody response). Data are shown for (A) Human respiratory syncytial virus Attachment Glycoprotein G (*G*), (B) Human adenovirus C penton protein (*L2*), and (C) Epstein-Barr virus nuclear antigen 1 (*EBNA1*). Data shown are the mean of two replicates.

majority of seropositive samples for a given virus recognized the same immunodominant peptides, suggesting that the antiviral B cell response is highly stereotyped across individuals. For example, in glycoprotein G from respiratory syncytial virus, there is only a single immunodominant peptide comprising positions 141-196 that is targeted by all samples with detectable antibodies to the protein, regardless of the country of origin (Figure 18A).

For other antigens, we observed inter-population serological differences. For example, two overlapping peptides from position 309-364 and 337-392 of the penton base protein from Adenovirus C frequently elicited antibody responses (Figure 18B). However, donors from the United States and South Africa had much stronger responses to peptide 309-364 ($p < 10^{-6}$, t-test) relative to donors from Thailand and Peru. We observed that for the EBNA1 protein from Epstein Barr virus, donors from all four countries frequently had strong responses to peptide 393-448 and occasionally to peptide 589-644. However, donors from Thailand and Peru had much stronger responses to peptide 57-112 ($p < 10^{-6}$, t-test) (Figure 18C). These differences may reflect variation in the strains endemic in each region. In addition, polymorphism of MHC class II alleles, immunoglobulin genes and other modifiers that shape immune responses in each population likely play a role in defining the relative immunodominance of antigenic peptides.

To determine whether the humoral responses that target an immunodominant peptide are actually targeting precisely the same epitope, we constructed single-, double-, and triple-alanine scanning mutagenesis libraries for 8 commonly recognized peptides. These were introduced into the same T7 bacteriophage display vector and subjected to the same immunoprecipitation and sequencing protocol using samples from the United States. Mutants that disrupt the epitope diminish antibody binding affinity and peptide enrichment. We found that for all 8 peptides

tested, there was a single, largely contiguous subsequence in which mutations disrupted binding for the majority of samples. As expected, the triple-mutants abolished antibody binding to a greater extent, and the enrichment patterns were similar among single-, double- and triple-mutants of the same peptide (Figure 19 through Figure 26). For 4 of the 8 peptides, a 9 to 15 amino acid region was critical for antibody recognition in >90% of samples (Figure 19 through Figure 22). One other peptide had a region of similar size that was critical in about half of the samples (Figure 23). In another peptide, a single region was important for antibody recognition in the majority of the samples, but the extents of the critical region varied slightly for different samples and occasionally there are donors that recognize a completely separate epitope (Figure 24). The remaining two peptides contained a single triple mutant that abolished binding in the majority of samples, but the critical region also extended further to different extents depending on the sample (Figure 25 and Figure 26). Surprisingly, in one of these peptides, in addition to the main region surrounding positions 13-14 that is critical for binding, a single G36A mutation disrupted binding in almost half of the samples whereas none of the double- or triple-alanine mutants that also included the adjacent positions (L35, G37) affected binding (Figure 26). It is possible that G36 plays a role in helping the peptide adopt an antigenic conformation and multiple-mutants containing the adjacent Leu or Gly residues rescue this ability. We occasionally saw other examples of mutations that resulted in patterns of disrupted binding with no simple explanation, illustrating the complexity of antibody-antigen interaction.

The discovery of recurring targeted epitopes led us to ask whether we could apply this knowledge to improve the sensitivity of viral detection with VirScan. We hypothesized that samples showing a strong response to a recurrently targeted “diagnostic” peptide, which we defined as a peptide enriched in at least 30% of known positive samples, are likely to be

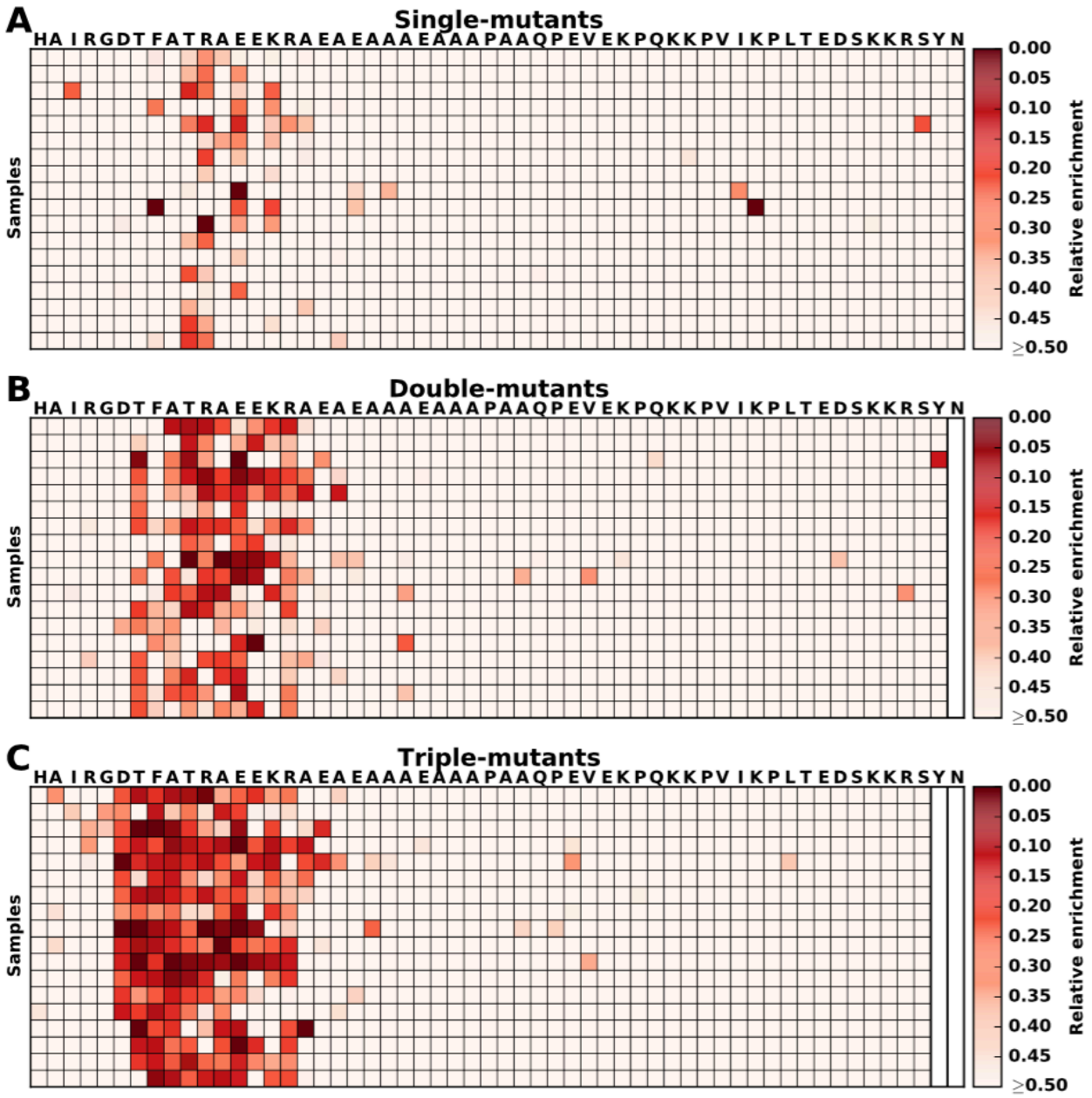

Figure 19. Recognition of common epitopes within an antigenic peptide from human adenovirus C penton protein (*L2*) across individuals. Each row is a sample. Each column denotes the first mutated position for the (A) single-, (B) double-, and (C) triple-alanine mutant peptide starting with the N-terminus on the left. Each double- and triple-alanine mutant contains two or three adjacent mutations, respectively, extending towards the C-terminus from the colored cell. The color intensity of each cell indicates the enrichment of the mutant peptide relative to the wild-type. For double-mutants, the last position is blank. The same is true for the last two positions for triple-mutants. Data shown are the mean of two replicates.

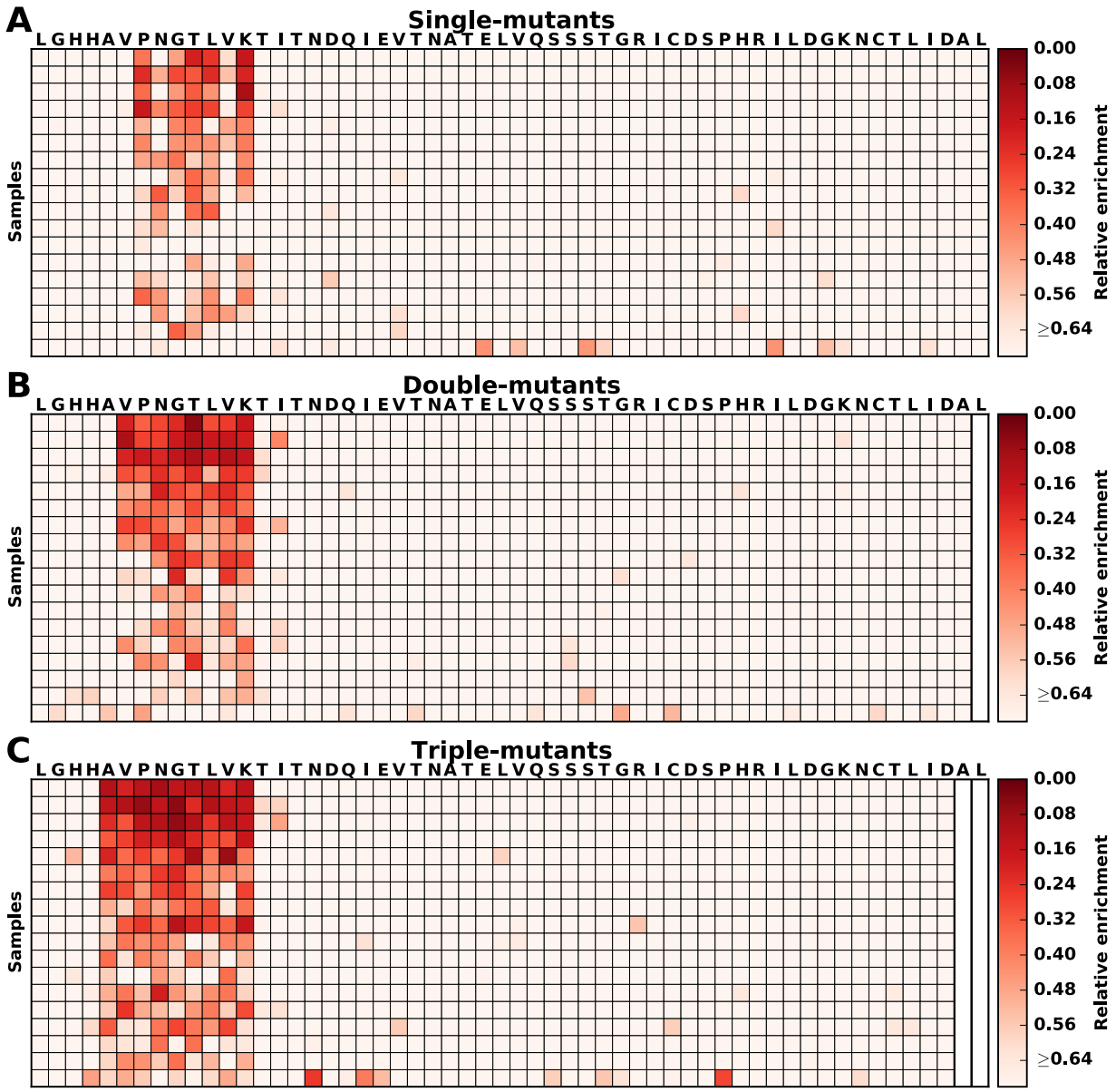

Figure 20. Scanning mutagenesis identification of linear B cell epitopes in an immunogenic peptide from Influenza A: hemagglutinin (UniProt ID: H8PET1, positions 1-56). Each row is a sample. Each column denotes the first mutated position for (a) single-, (b) double-, and (c) triple-alanine mutant peptides. The color intensity of each cell indicates the enrichment of the mutant peptide relative to the wild-type. For double-mutants, the last position is blank. The same is true for the last two positions for triple-mutants. Data shown are the mean of two replicates

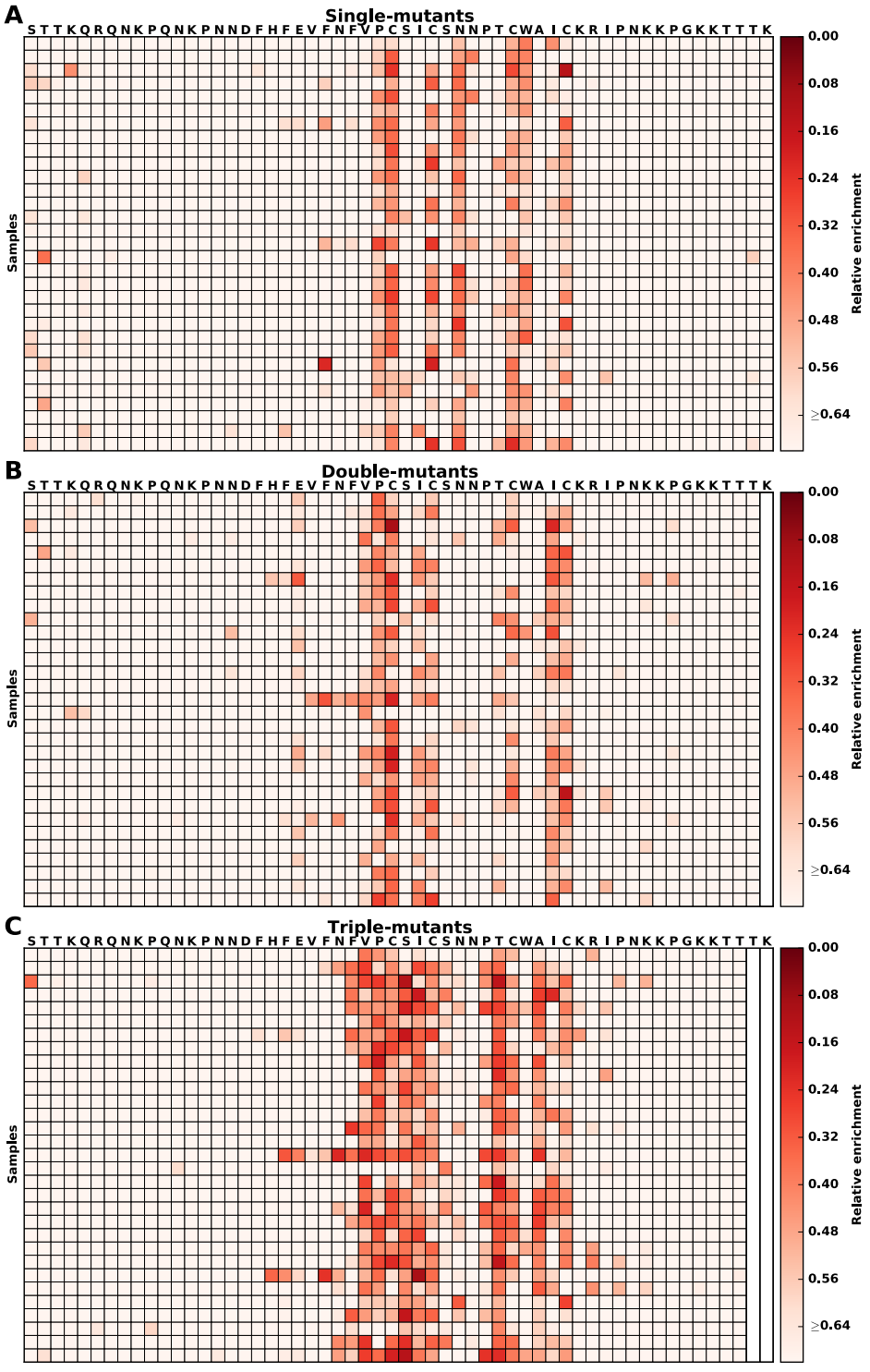

Figure 21. See caption for Figure 20. Respiratory syncytial virus: attachment G glycoprotein (UniProt ID: P03276, positions 337-392)

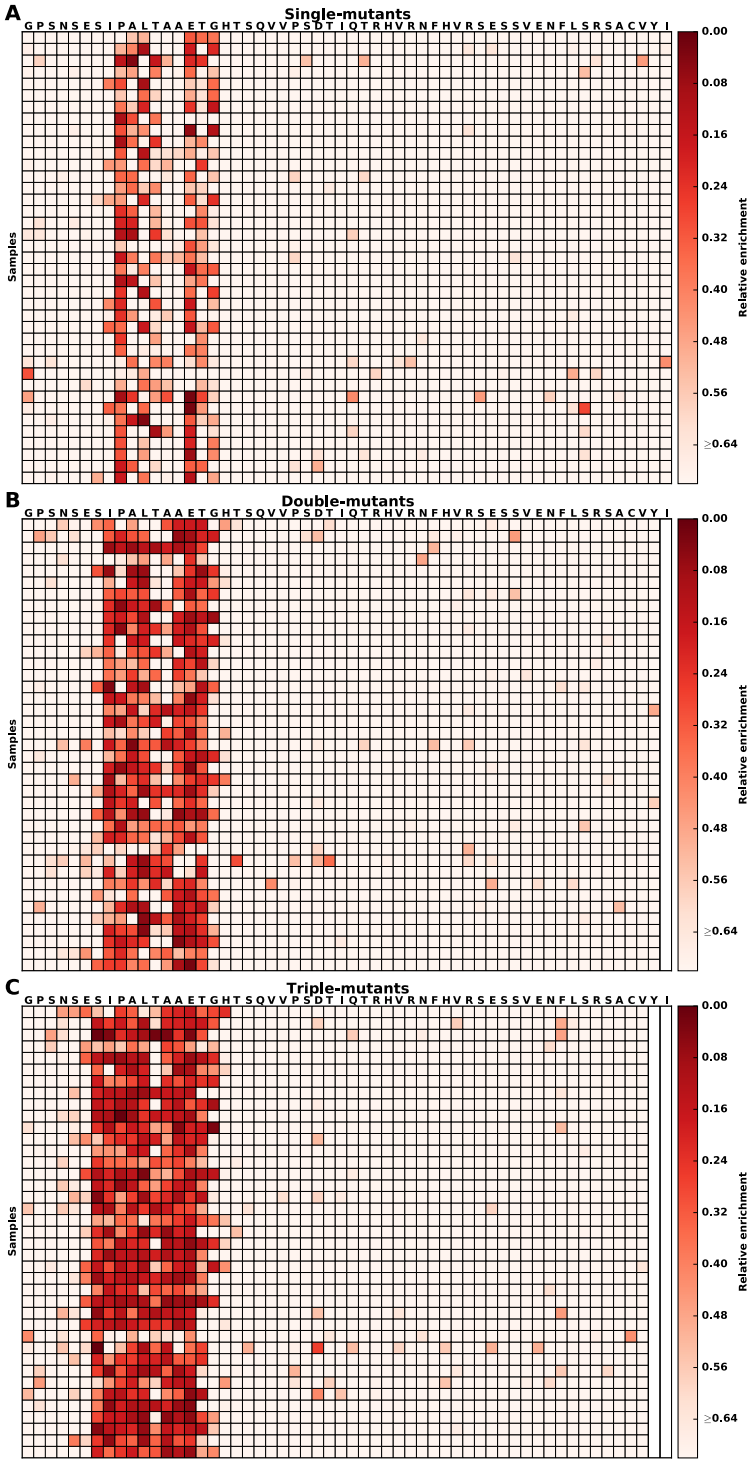

Figure 22. See caption for Figure 20. Enterovirus B: genome polyprotein (UniProt ID: Q66474, positions 561-616).

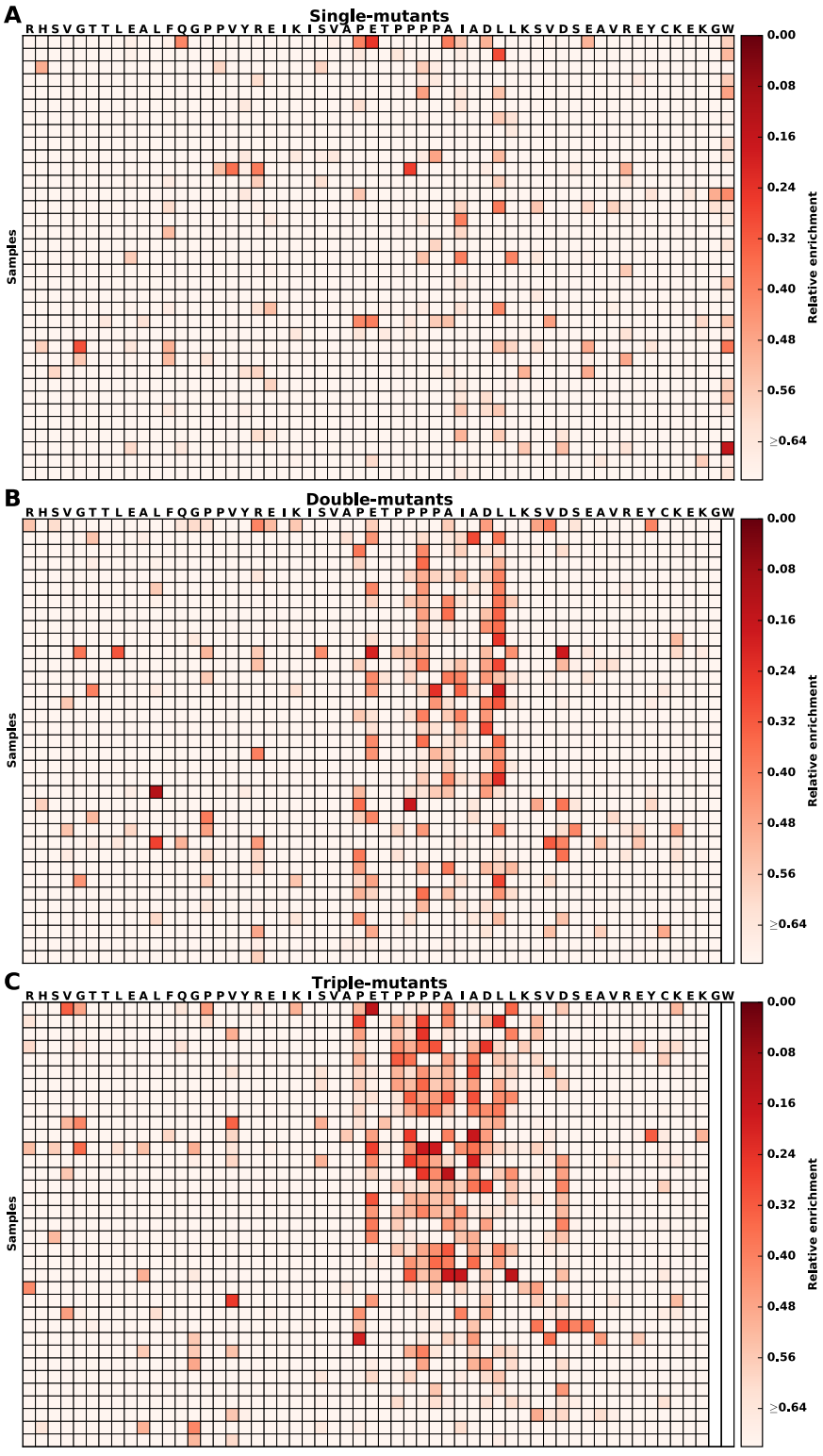

Figure 23. See caption for Figure 20. Enterovirus B: genome polyprotein (UniProt ID: Q6W9F9, positions 1429-1484).

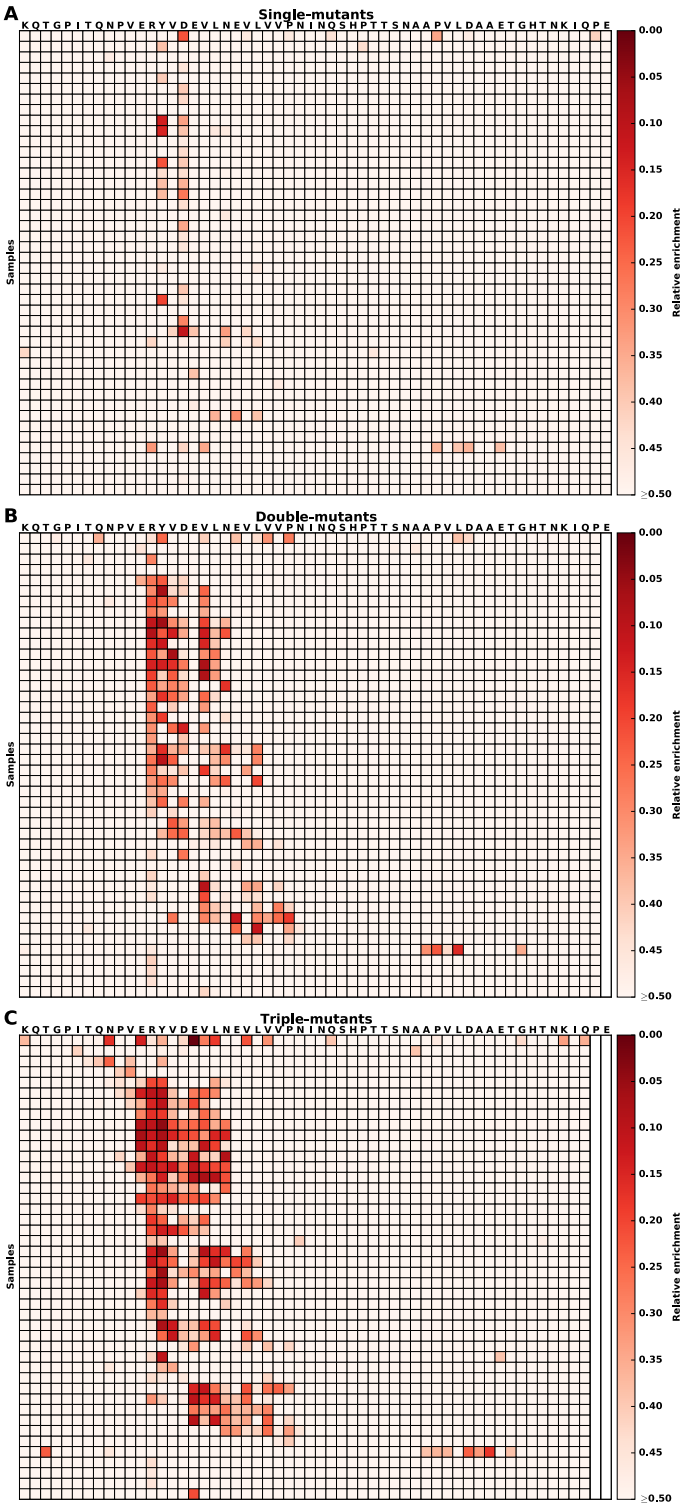

Figure 24. See caption for Figure 20. Rhinovirus A: genome polyprotein (UniProt ID: Q82122, positions 561-616)

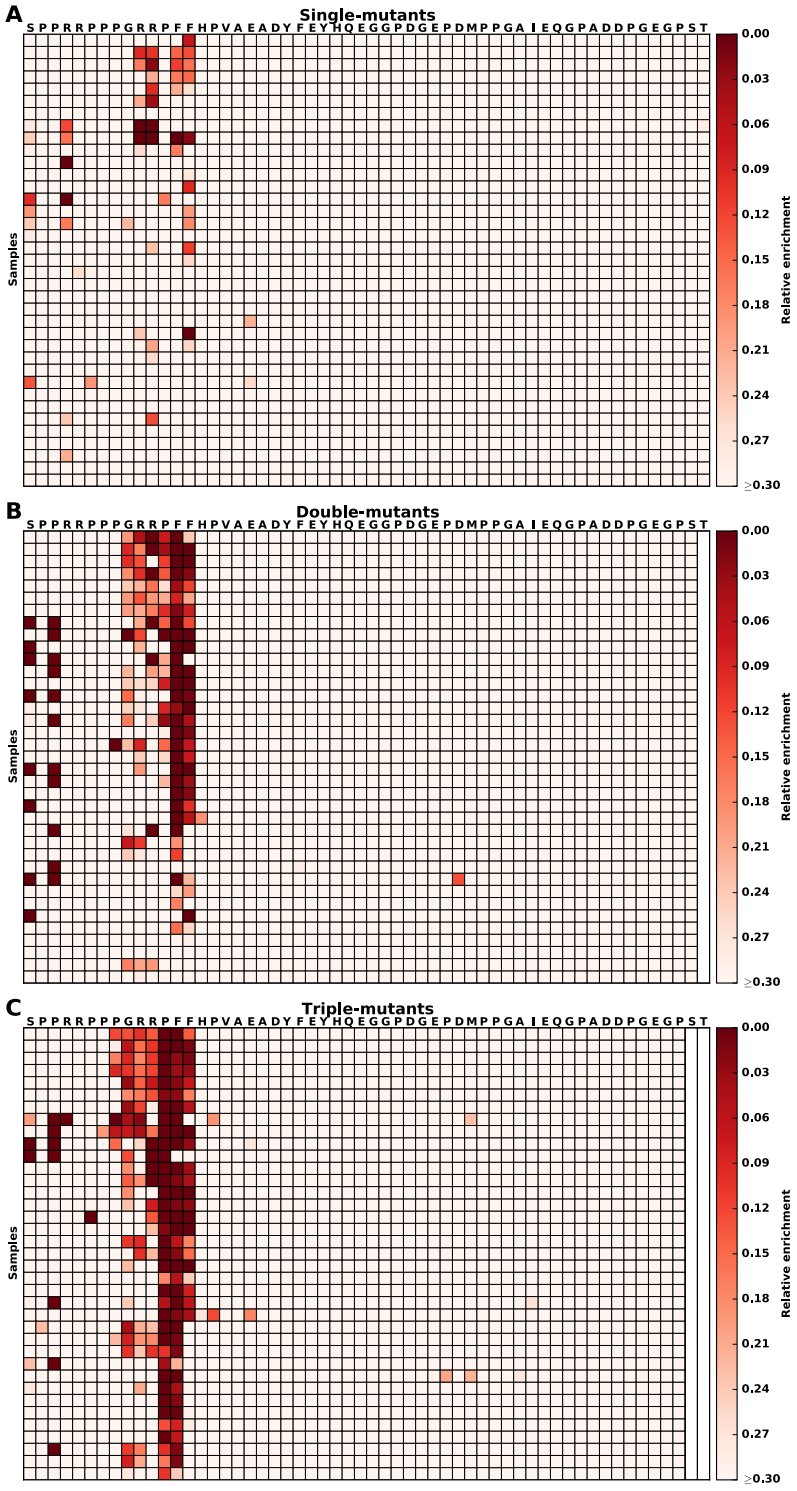

Figure 25. See caption for Figure 20. Epstein-Barr virus: nuclear antigen 1(UniProt ID: Q1HVF7, positions 393-448).

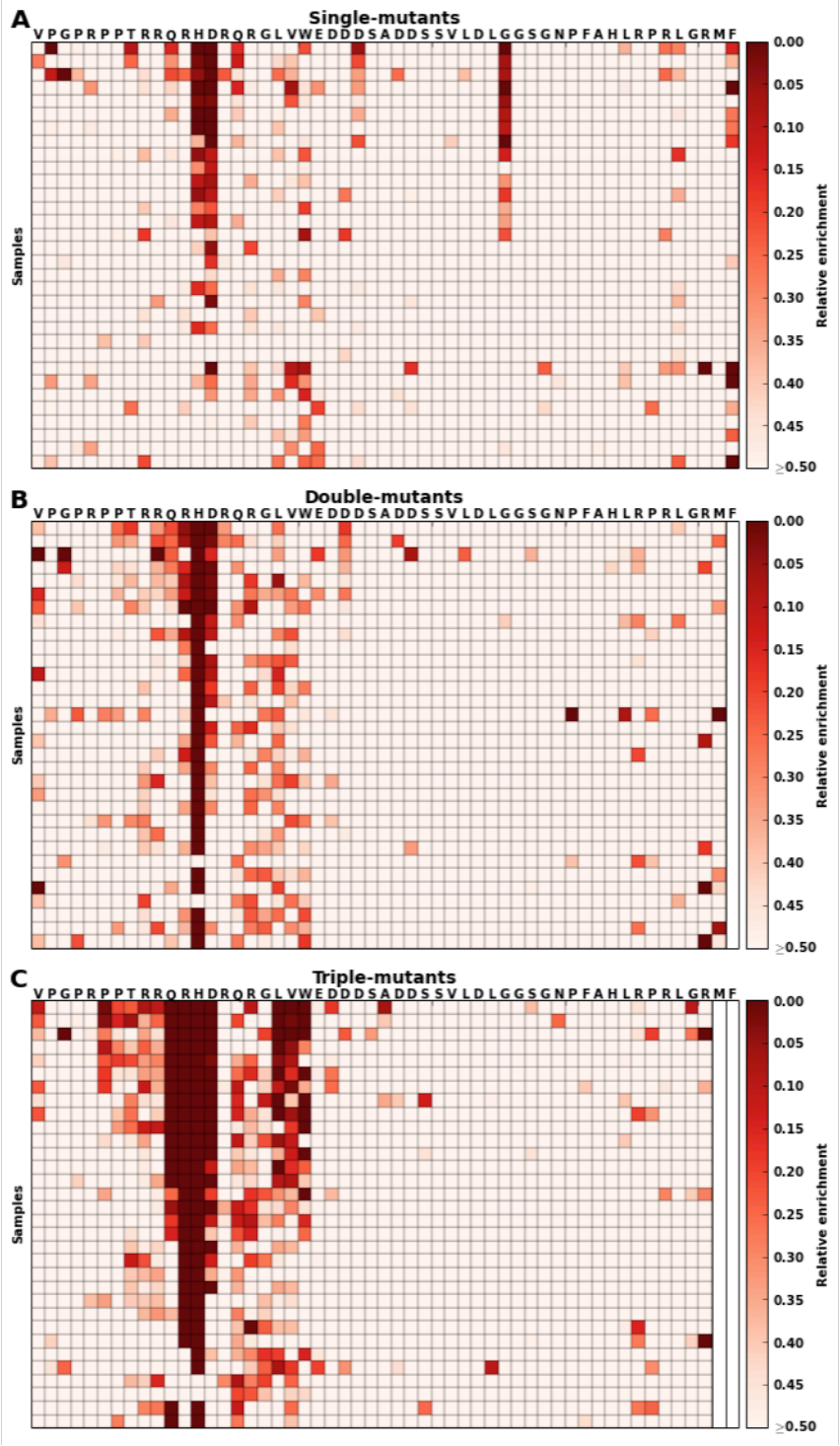

Figure 26. See caption for Figure 20. Adenovirus C: precapsid vertex protein (UniProt ID: P03279, positions 533-585)

seropositive even if they do not meet our stringent cutoff requiring at least two non-overlapping enriched peptides. We tested how this modified criterion affected our sensitivity and specificity in detecting HIV and HCV and found that it reduced the number of false negatives without affecting the specificity of the assay (fig. S13). We next turned our attention to respiratory syncytial virus (RSV), a virus for which our detected seroprevalence was lower than reported epidemiological rates, suggesting imperfect sensitivity of our assay. We tested 60 patient sera for antibodies to RSV by ELISA and found 95% were positive, above the reported sensitivity of the assay and consistent with near-universal exposure to this pathogen. Applying the modified criterion to these samples increased our rate of detection by VirScan from 63% to 97% (Table 4). These data suggest that assigning more weight to recurrently targeted epitopes can enhance the sensitivity of VirScan and that the performance of the assay can be improved by screening known positives for a particular virus.

Initial VirScan algorithm		**RSV ELISA**	
		Positive	**Negative**
RSV VirScan	**Positive**	37	1
	Negative	20	2

With diagnostic peptides		**RSV ELISA**	
		Positive	**Negative**
RSV VirScan	**Positive**	55	3
	Negative	2	0

Table 4. Modified algorithm applying more weight to diagnostic peptides shows improved detection of antibodies against RSV. 60 patient sera were screened for RSV antibodies by ELISA and with VirScan. The concordance of the ELISA results with the initial and modified VirScan algorithms is shown in the tables.

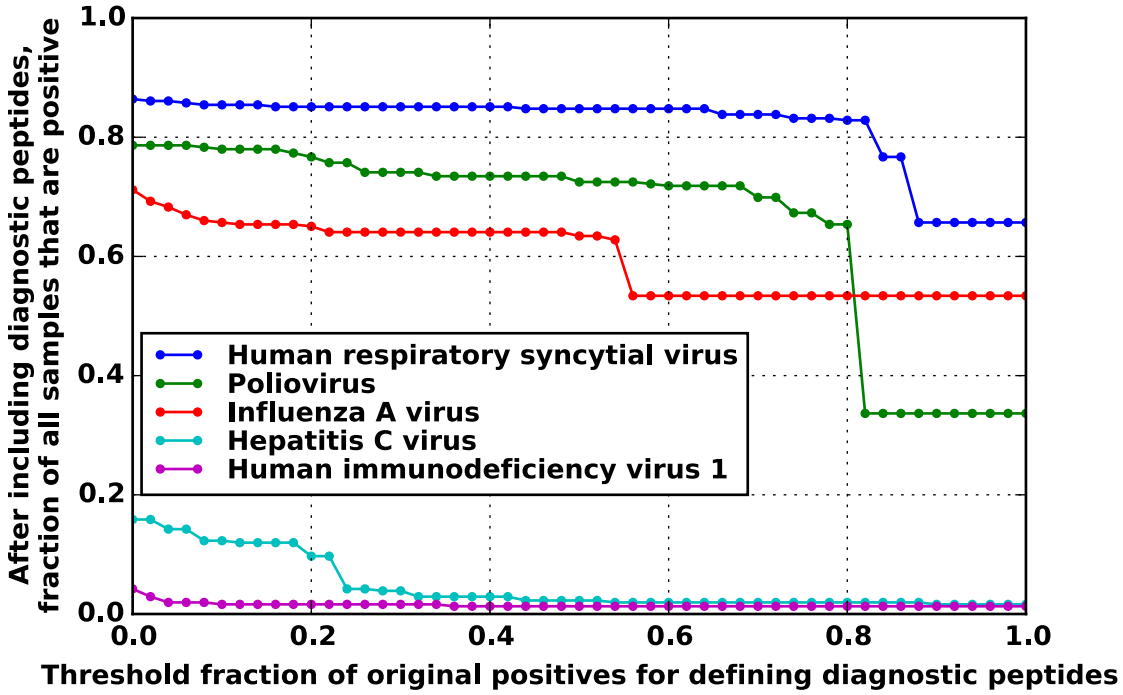

Figure 27. Increased sensitivity after including samples targeting "diagnostic" peptides. For each virus, we examined all the samples that enriched multiple peptides that share a single epitope. If this epitope is "diagnostic" (i.e., recurrently targeted in at least a threshold fraction of the samples that were originally called positive for that virus), we considered the sample to be positive for that virus. The y-axis shows the fraction of samples that are considered positive after including these samples. The x-axis represents the minimum fraction of the original positive samples that must enrich a peptide for it to be considered diagnostic. Using a threshold of 30-70% significantly increases the rate of detecting respiratory syncytial virus without significantly increasing the rate of detecting hepatitis C and HIV, which should have low seroprevalence in this population (only samples from the United States that were not known HIV or HCV positives were included in this analysis).

Discussion

We have developed VirScan, a technology for identifying viral exposure and B cell epitopes across the entire known human virome in a single, multiplex reaction using less than a drop of blood. VirScan uses DNA microarray synthesis and bacteriophage display to create a uniform, synthetic representation of peptide epitopes comprising the human virome.

Immunoprecipitation and high-throughput DNA sequencing reveals the peptides recognized by antibodies in the sample. VirScan is easily automated in 96-well format to enable high throughput sample processing. Barcoding of samples during PCR enables pooled analysis that can dramatically reduce the per-sample cost. The VirScan approach has several advantages for studying the effect of viruses on the host immune system. By detecting antibody responses, it can identify infectious agents that have been cleared after an effective host response. Current serological methods of antiviral antibody detection typically employ the selection of a single optimized antigen in order to achieve high accuracy. In contrast, VirScan's unique approach does not require such optimization in order to obtain similar performance. VirScan achieves sensitive detection by assaying each virus's complete proteome to detect any antibodies directed to epitopes that can be captured in a 56-residue fragment and specificity by computationally eliminating cross-reactive antibodies. This unbiased approach identifies exposure to less well-studied viruses for which optimal serological antigens are not known and can be rapidly extended to include new viruses as they are discovered (*22*).

While sensitive and selective, VirScan also has a few limitations. First, it cannot detect epitopes that require post-translational modifications. Secondly, it cannot detect epitopes that involve discontinuous sequences on protein fragments greater than 56 residues. In principle, the latter can be overcome by using alternative technologies that allow for the display of full-length proteins such as Parallel Analysis of Translated ORFs (PLATO) (*85*). Third, VirScan is likely to be less specific compared with certain nucleic acid tests that discern highly related virus strains. However, VirScan demonstrates excellent serological discrimination among similar virus species, such as HSV1 and HSV2 and can even distinguish the genotype of HCV 69% of the time. We envision VirScan will become an important tool for first-pass unbiased serologic

screening applications. Individual viruses or viral proteins uncovered in this way can subsequently be analyzed in further detail using more focused assays, as we have demonstrated for a panel of immunodominant epitopes.

We have demonstrated that VirScan is a sensitive and specific assay for detecting exposure to viruses across the human virome. Because it can be performed in high-throughput and requires minimal sample and cost, VirScan enables rapid and cost-effective screening of large numbers of samples to identify population-level differences in virus exposure across the human virome. In this work, we analyzed over 106 million antibody-viral peptide interactions in a comprehensive study of pan-virus serology in a large, diverse population. In doing so we detected 84 different viral species in 2 or more individuals. This is likely to be an underestimate of the history of viral infection as only low levels of circulating antibodies may remain from infections that were cleared in the distant past. In addition, an individual could be infected by multiple distinct strains of each viral species. We identified known and novel differences in virus exposure between groups differing in age, HIV status, and geographic location across four different continents. Our results are largely consistent with previous studies, validating the effectiveness of VirScan. For example, cytomegalovirus antibodies were found in significantly higher frequencies in Peru, Thailand, and South Africa whereas Kaposi's sarcoma-associated herpesvirus and HSV1 antibodies were detected more frequently in Peru and South Africa, but not in Thailand (*76*, *86–90*). We also uncovered previously undocumented serological differences, such as an increased rate of antibodies against Adenovirus B and respiratory syncytial virus in HIV positive individuals compared to HIV negative individuals. These differences may provide insight into how HIV co-infection alters the balance between host immunity and resident viruses, as well as help to identify pathogens that may increase

susceptibility to HIV and other heterologous infections. HIV infection may reduce the immune system's ability to control reactivation of normally dormant resident viruses or to prevent opportunistic infections from taking hold and triggering a strong adaptive immune response. Beyond the epidemiological applications demonstrated here, VirScan could also be applied to identify viral exposures that correlate with disease or other phenotypes in virome-wide association studies.

Our results identified a large number of novel B cell epitopes, cumulatively nearly doubling the number of all previously identified viral epitopes. We have utilized our data to identify globally immunodominant and commonly recognized "public" epitopes. For most species of viruses, one or more peptides are individually recognized in over 70% to 95% of samples positive for that species (table S3). We identified a set of two peptides that together are recognized by >95% of all screened samples and a set of five peptides that together are recognized in >99% of screened samples.

These public epitopes could be used to improve vaccine design by piggybacking on the existing antibody response against them. Fusing a public B cell epitope to a protein in a vaccine to which we hope to induce an immune response may increase a vaccine's efficacy among a broad population by improving presentation of that protein and aiding affinity maturation. Pre-existing B cells recognizing the public epitope can act as antigen presenting cells to process and present T cell epitopes of the fused vaccine target on MHC class I and II (*91*). Antibodies secreted by these B cells can also participate in immune complexes with the fused vaccine target, which are critical for follicular dendritic cells to prime class switching and affinity maturation of B cells recognizing other epitopes on the fused antigen (*92*). Finally, we demonstrated that

Species	Protein	Peptide	%
Rhinovirus B	Genome polyprotein	QTVALTEGLGDELEEVIVEKTKQTVASI SSGPKHTQKVPILTANETGATMPVLPSD	95%
Human herpesvirus 5	Envelope glycoprotein M	TASGEEVAVLSHHDSLESRRLREEEDDD DDEDFEDA	90%
Enterovirus B	Genome polyprotein	PFIQQEAKLQGEPGKAIESAISRVADTISS GPTNSEQVPALTAAETGHTSQVVPGD	86%
Human respiratory syncytial virus	Attachment glycoprotein	NKPSTKPRPKNPPKKPKDDYHFEVFNFV PCSICGNNQLCKSICKTIPSNKPKKKPT	85%
Human herpesvirus 4	Epstein-Barr nuclear antigen 1	SPPRRPPPGRRPFFHPVAEADYFEYHQEG GPDGEPDMPPGAIEQGPADDPGEGPST	81%
Human herpesvirus 1	Envelope glycoprotein D	RRHTQKAPKRIRLPHIREDDQPSSHQPLFY	80%
Norwalk virus	Genome polyprotein	LSSMAVTFKRALGGRAKQPPPRETPQRPP RPPTPELVKKIPPPPPNGEDELVVSYS	77%
Human adenovirus C	Pre-histone-like nucleoprotein	MTQGRRGNVYWVRDSVSGLRVPVRTRP PRN	74%
Enterovirus C	Genome polyprotein	QGALTLSLPKQQDSLPDTKASGPAHSKE VPALTAVETGATNPLAPSDTVQTRHVVQ	73%
Human herpesvirus 3	Envelope glycoprotein C	PDPAVAPTSAASRKPDPAVAPTSAASRK PDPAVAPTSAATRKPDPAVAPTSAASRK	72%
Human immunodeficiency virus 1	Envelope glycoprotein gp160	ERYLKDQQLLGIWGCSGKLICTTAVPWNA SWSNKSLEQIWNNMTWMEWDREINNYT	60%

Table 5. Certain peptides are commonly targeted by the antibody response. We determined the peptide from each species of virus that was most frequently targeted in donors that were exposed to that virus. In each row, the frequency is the percentage of samples positive for the species of virus that had antibodies targeting the peptide sequence shown. The parent protein of the peptide is also listed.

applying more weight to these public epitopes increases the sensitivity of VirScan without significantly affecting specificity, suggesting that this limited subset of peptides can serve as the basis for the next generation of our assay or for other novel diagnostics.

We also found that the precise epitopes recognized by the B cell response are highly similar among individuals across many viral proteins. One possible model for this striking similarity is that these regions possess properties favorable for antigenicity, such as accessibility. Another model is that the same or highly similar B cell receptor sequences that recognize these epitopes are commonly generated. Identical T cell receptor sequences (“public” clonotypes) have been found in multiple individuals and are thought to be the result of biases during the

recombination process that favor certain amino acid sequences (*93*). V(D)J recombination of the immunoglobulin heavy and light chain loci is also heavily biased (*94*). Highly similar or even identical complementarity determining region 3 (CDR3) sequences have been observed in dengue virus specific antibodies from different individuals (*95*). It is possible that, rather than being an exception for dengue specific antibodies, this represents a general phenomenon: inherent biases in V(D)J recombination generate the same or similar antibodies in multiple individuals that recognize highly similar epitopes. Slight differences in the antibody CDR3 sequence may subtly alter antibody-antigen interaction, leading to the slight variations observed in the extent of critical epitope regions. Sequencing of antigen specific antibody genes will be required to investigate these possibilities. The same principle may also apply to T cell epitopes and their cognate TCRs.

In conclusion, VirScan is a new method that enables human virome-wide exploration - at the epitope level - of immune responses in large numbers of individuals. We have demonstrated its effectiveness for determining viral exposure and characterizing viral B cell epitopes in high throughput and at high resolution. Our preliminary studies have revealed intriguing general properties of the human immune system, both at the individual and population scale. VirScan will be an important tool in uncovering the effect of host-virome interactions on human health and disease and could easily be expanded to include other human pathogens such as bacteria, fungi and protozoa.

Supplementary Discussion

Estimating VirScan's specificity

Although we detected antibody responses to rare and highly virulent viruses such as Marburg and bat lyssavirus, they were found in less than 1% of the population (Table 3), indicating that specificity is over 99% for these viruses, which is similar to the results in Table 2. Because we screened hundreds of sera for recognition of 206 virus species each, we performed the equivalent of approximately 100,000 individual tests, and eliminating such false positives altogether would require specificity of approximately 99.999% for each virus. Even with 99% specificity, a test will have 1% false positives, or approximately three per virus species for the 303 samples in population analyzed in Table 3.

In addition, 92 species of virus out of 206 were not detected in any samples from this population. Another 45 were detected in 3 or fewer samples. Assuming these are all false positives, which errs on the side of overestimating false positives, this analysis suggests that the specificity is 99.9%. While this is an imperfect estimate because we do not know how many of the detected positives are actually false positives, it gives an approximate estimate that argues the specificity is very high. No assay is perfect, and even highly optimized ELISAs for single viruses have some level of false positive, but our results give us a great deal of confidence in VirScan's specificity.

Differentially weighting recurrent peptides increases sensitivity

After discovering that certain epitopes are recurrently targeted, we examined whether we could apply this knowledge to improve the sensitivity of viral detection with VirScan. Recurrent epitopes make up a very small portion of a virus's proteome. On average, less than 1% of a given virus's proteome is targeted in more than 30% of samples positive for that virus. We hypothesized that samples showing a strong response to these recurrently targeted "diagnostic" peptides, which we defined as a peptide enriched in at least 30% of positive samples, are likely

to be seropositive even if they do not meet our stringent cutoff requiring at least two non-overlapping enriched peptides. Thus, we introduced a modified criterion for calling a sample positive for a given virus that only requires one unique enriched peptide from the virus as long as the peptide is diagnostic (i.e., enriched in at least 30% of the samples that were originally called positive for that virus) and at least one other peptide that shares at least 7aa sequence homology was also enriched. The requirement for enrichment of two or more related peptides guards against potentially spurious technical enrichments.

We tested how this modified criterion affected our sensitivity and specificity in the known HCV positive and negative samples. In this set of samples, we had two false negatives, which had 11 and 14 enriched peptides, respectively, that were highly homologous and thus filtered down to one unique epitope. In both samples, this epitope corresponded to the N-terminus of the genome polyprotein, which is targeted in over 70% of the HCV positive samples. Thus, the modified criterion increases sensitivity for HCV to 100%. This modified criterion does not lead to increased false positives among known HCV negative samples nor does it significantly increase the rate of detecting HCV positive samples among the rest of the US samples (Figure 27).

We then tested how this modified criterion works on the known HIV positive and negative samples. Of the four false negative samples, one had enriched six related peptides targeted in 70-90% of the HIV positive samples and would be considered positive using the modified criterion. This relaxed criterion does not lead to increased false positives among known HIV negative samples nor does it increase the rate of detecting HIV positive samples among the rest of the US samples (Figure 27). The remaining three false negative samples did not

significantly enrich a recurrently targeted peptide. However, upon further examination, we found that in two of the samples, although no single recurrent peptide was enriched, the set of recurrently targeted peptides were, as a group, enriched relative to other HIV peptides using a modified Gene Set Enrichment Analysis approach (Figure 28). These results suggest that these false negatives are due to low titers of anti-HIV antibodies that do not pass our stringent threshold for significance for any one peptide, but are significant when the set of homologous peptides are considered together. Once recurring peptides are identified for a given virus, this methodology could be used to develop a secondary analysis criterion for suspected false negatives, especially those that present with some but too few scoring peptides to meet the threshold for consideration as a positive.

We next turned our attention to respiratory syncytial virus (RSV), a virus for which our detected seroprevalence was lower than reported epidemiological rates, suggesting imperfect sensitivity of our assay. We tested 60 patient sera for antibodies to RSV by ELISA and found 95% were positive, above the reported sensitivity of the assay and consistent with near-universal exposure to this pathogen. Applying the modified criterion to these samples increased our rate of detection by VirScan from 63% to 97% (Table 4). These data suggest that assigning more weight to recurrently targeted epitopes can enhance the sensitivity of VirScan and that the performance of the assay can be improved by screening known positives for a particular virus to discover these recurrently targeted epitopes.

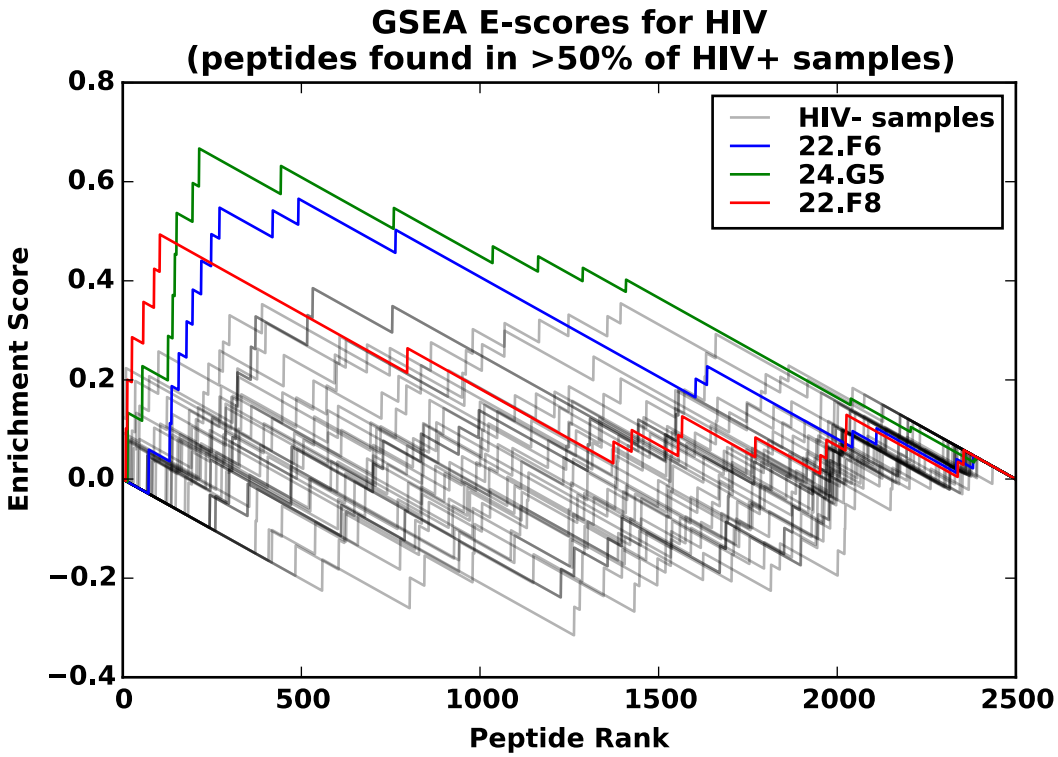

Figure 28. Peptide set enrichment analysis for peptides containing recurrent epitopes in HIV samples. The analysis and graph are similar to the enrichment score calculation for the Gene Set Enrichment Analysis method. For each sample, the HIV peptides that did not pass our threshold for significantly enriched were ranked in descending order of $-\log_{10}$(p-value). A running sum was calculated by going down the list and, if the peptide was recurrently targeted in HIV (enriched in the majority of the HIV positive samples), the running sum was incremented by a value weighted by the $-\log_{10}$(p-value) of the peptide and normalized to 1 for all recurrent peptides. Otherwise, the running sum was decrement by a fixed value that was normalized to 1 for all non-recurrent peptides. The running sum is plotted for the 31 HIV negative samples (black lines) and for the HIV false negative samples (blue, green, and red lines). The maximum positive displacement of the running sum (enrichment score) is a measure of how significantly the set of peptides is enriched relative to the other HIV peptides.

Methods

Design and cloning of viral peptide and scanning mutagenesis library sequences

For the virome peptide library, we first downloaded all protein sequences in the UniProt database from viruses with human host and collapsed on 90% sequence identity (http://www.uniprot.org/uniref/?query=uniprot:(host: "Human+[9606]")+identity:0.9). The clustering algorithm UniProt represents each group of protein sequences sharing at least 90% sequence similarity with a single representative sequence. Then, we created 56 aa peptide sequences tiling through all the proteins with 28 aa overlap. We reverse translated these peptide sequences into DNA codons optimized for expression in *E. coli*, making synonymous mutations when necessary to avoid restriction sites used in subsequent cloning steps (EcoRI and XhoI). Finally, we added the adapter sequence "aGGAATTCCGCTGCGT" to the 5' end and "CAGGgaagagctcgaa" to the 3' end to form the 200 nt oligonucleotide sequences.

For the scanning mutagenesis library, we first took the sequences of the peptides to be mutagenized. For each peptide, we made all single-mutants, and consecutive double- and triple-mutants sequences scanning through the whole peptide. Non-alanine amino acids were mutated to alanine and alanines were mutated to glycine. We reverse translated these peptide sequences into DNA codons, making synonymous mutations when necessary to avoid restriction sites used in subsequent cloning steps (EcoRI and XhoI). We also made synonymous mutations to ensure that the 50 nt at the 5' end of peptide sequence is unique to allow unambiguous mapping of the sequencing results. Finally, we added the adapter sequence "aGGAATTCCGCTGCGT" to the 5' end and "CAGGgaagagctcgaa" to the 3' end to form the 200 nt oligonucleotide sequences.

The 200 nt oligonucleotide sequences were synthesized on a releasable DNA microarray. We PCR amplified the DNA using the primers T7-PFA (aatgatacggcggGAATTCCGCTGCGT) and T7-PRA (caagcagaagACTCGAGCTCTTCCCTG), digested the product with EcoRI and

XhoI, and cloned the fragment into the EcoRI/SalI site of the T7FNS2 vector (*19*). The resulting library was packaged into T7 bacteriophage using the T7 Select Packaging Kit (EMD Millipore) and amplified using the manufacturer suggested protocol.

Phage immunoprecipitation and sequencing

We performed phage immunoprecipitation and sequencing using a slightly modified version of previously published PhIP-Seq protocols (*19*, *70*). First, we blocked each well of a 96 deep-well plate with 1 mL of 3% BSA in TBST overnight on a rotator at 4ºC. To each pre-blocked well, we added sera or plasma containing approximately 2 µg of IgG (quantified using a Human IgG ELISA Quantitation Set (Bethyl Laboratories)) and 1 mL of the bacteriophage library diluted to approximately 2×10^5 fold representation (2×10^{10} pfu for a library of 10^5 clones) in phage extraction buffer (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 6 mM $MgSO_4$). We performed two technical replicates for each sample. We allowed the antibodies to bind the phage overnight on a rotator at 4ºC. The next day, we added 20 µL each of magnetic Protein A and Protein G Dynabeads (Invitrogen) to each well and allowed immunoprecipitation to occur for 4 h on a rotator at 4ºC. Using a 96-well magnetic stand, we then washed the beads three times with 400 µL of PhIP-Seq wash buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.1% NP-40). After the final wash, we resuspended the beads in 40 µL of water and lysed the phage at 95 ºC for 10 m. We also lysed phage from the library before immunoprecipitation ("input") and after immunoprecipitation with beads alone.

We prepared the DNA for multiplexed Illumina sequencing using a slightly modified version of a previously published protocol (*96*). We performed two rounds of PCR amplification on the lysed phage material using hot start Q5 polymerase according to the manufacturer suggested protocol (NEB). The first round of PCR used the primers IS7_HsORF5_2

(ACACTCTTTCCCTACACGACTCCAGTCAGGTGTGATGCTC) and IS8_HsORF3_2 (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCGAGCTTATCGTCGTCATCC). The second round of PCR used 1 μL of the first round product and the primers IS4_HsORF5_2 (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACTCCAGT) and a different unique indexing primer for each sample to be multiplexed for sequencing (CAAGCAGAAGACGGCATACGAGATxxxxxxxGTGACTGGAGTTCAGACGTGT, where "xxxxxxx" denotes a unique 7 nt indexing sequence). After the second round of PCR, we determined the DNA concentration of each sample by qPCR and pooled equimolar amounts of all samples for gel extraction. Following gel extraction, the pooled DNA was sequenced by the Harvard Medical School Biopolymers Facility using a 50 bp read cycle on an Illumina HiSeq 2000 or 2500. We pooled up to 192 samples for sequencing on each lane and generally obtained approximately 100 - 200 million reads per lane (500,000 to 1,000,000 reads per sample).

Informatics and statistical analysis

We performed the initial informatics and statistical analysis using a slightly modified version of the previously published technique (*19*, *70*). We first mapped the sequencing reads to the original library sequences using Bowtie and counted the frequency of each clone in the "input" and each sample "output" (*62*). Since the majority of clones are not enriched we use the observed distribution of output counts as a null distribution. We found that a zero-inflated generalized poisson distribution fits our output counts well. We use this null distribution to calculate a p-value for the likelihood of enrichment for each clone. The probability mass function for the zero-inflated generalized poisson distribution is

$$P(Y = y) = \begin{cases} \pi + (1-\pi)(\theta(\theta+\lambda)^{x-1}e^{-\theta-x\lambda}), & y = 0 \\ (1-\pi)(\theta(\theta+\lambda)^{x-1}e^{-\theta-x\lambda}), & y > 0 \end{cases}$$

We used maximum likelihood estimation to regress the parameters π, θ, and λ to fit the distribution of counts after immunoprecipitation for all clones present at a particular frequency count in the input. We repeated this procedure for all of the observed input counts and found that θ and λ are well fit by linear regression and π by an exponential regression as a function of input count (Figure 9). Finally, for each clone we used its input count and the regression results to determine the null distribution based on the zero-inflated generalized poisson model, which we used to calculate the $-\log_{10}$(p-value) of obtaining the observed count.

To call hits, we determined the threshold for reproducibility between technical replicates based on a previously published method (*70*). Briefly, we made scatter plots of the log10 of the -log10 (p-values) and used a sliding window of width 0.005 from 0 to 2 across the axis of one replicate. For all the clones that fell within each window, we calculated the median and median absolute deviation of the log10 of the –log10 (p-values) in the other replicate and plotted it against the window location (Figure 10). We called the threshold for reproducibility the first window in which the median was greater than the median absolute deviation. We found that the distribution of the threshold –log10 (p-value) was centered around a mean of approximately 2.3 (Figure 14). So we called a peptide a hit if the –log10 (p-value) was at least 2.3 in both replicates. We eliminated the 593 hits that came up in at least three of the twenty-two immunoprecipitations with beads alone (negative control for non-specific binding). We also filtered out any peptides that were not enriched in at least two of the samples.

To call virus exposures, we grouped peptides according to the virus the peptide is derived from. We grouped all peptides from individual viral strains for which we had complete proteomes. The sample was counted as positive for a species if it was positive for any strain from

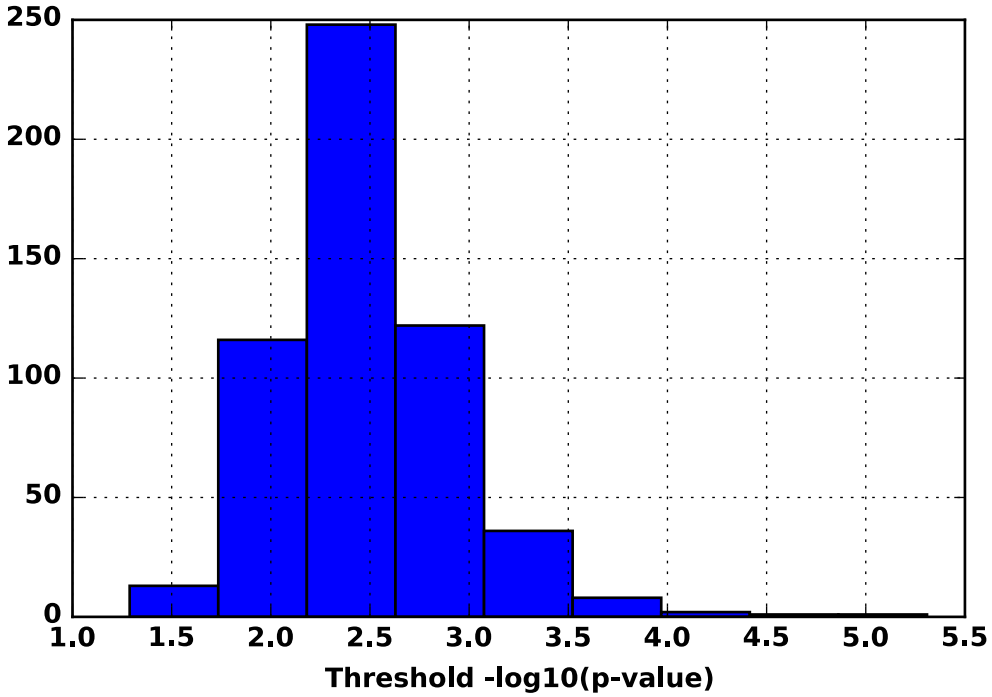

Figure 29. Distribution of reproducibility threshold -log10(p-values). Histogram of the frequency of the reproducibility threshold -log10(p-values). The mean and median of the distribution are both approximately 2.3.

that species. For viral strains that had partial proteomes, we grouped them with other strains from the same species to form a complete set and bioinformatically eliminated homologous peptides (see next paragraph). We set a threshold number of hits per virus based on the size of the virus. We found that there is approximately a power-law relationship between size of the virus and the average number of hits per sample (Figure 12). In comparing results from VirScan to samples with known infection, we empirically determined that a threshold of 3 hits for herpes simplex virus 1 worked the best. We used this value and the slope of the best fit line to scale the threshold for other viruses. We also set a minimum threshold of at least 2 hits in order to avoid false positives from single spurious hits.

To bioinformatically remove cross-reactive antibodies, we first sorted the viruses by total number of hits in descending order. We then iterated through each virus in this order. For each

virus, we iterated through each peptide hit. If the hit shared a subsequence of at least 7 aa with any hit previously observed in any of the viruses from that sample, that hit was considered to be from a cross-reactive antibody and would be ignored for that virus. Otherwise, the hit is considered to be specific and the score for that virus is incremented by one. In this way, we summed only the peptide hits that do not share any linear epitopes. We compared the final score for each virus to the threshold for that virus to determine whether the sample is positive for exposure to that virus.

To identify differences between populations, we first used Fisher's exact test to calculate a p-value for the significance of association of virus exposure with one population versus another. Then, we constructed a null-distribution of Fisher's exact p-values by randomly permuting the sample labels 1000 times and re-calculating the Fisher's exact p-value for each virus. Using this null-distribution, we calculated the false discovery rate by dividing the number of permutation p-values more extreme than the one observed by the total number of permutations.

IEDB epitope overlap analysis

We downloaded data for all continuous human B cell epitopes from IEDB and filtered out all non-viral epitopes (*83*). To avoid redundancy in these 4,549 viral epitopes, we grouped together epitopes that are 100% identical or share a 7 aa subsequence, giving us 1,559 non-redundant epitope groups. Of these groups, 1,392 contain a member epitope that is also a subsequence of a peptide in the VirScan library. This represents the total number of epitopes we could detect by VirScan. To determine the number of epitopes we detected, we tallied the number of epitope groups with at least one member that is contained in a peptide that was enriched in one or two samples. Finally, to determine the number of non-redundant new epitopes

we detected, we grouped non-IEDB epitopes containing peptides that share a 7 residues subsequence and counted the number of these non-redundant peptide groups.

Scanning mutagenesis data analysis

First, we estimated the fractional abundance of each peptide by dividing the number of reads for that peptide by the total number of reads for the sample. Then, we divided the fractional abundance of each peptide after immunoprecipitation by the fractional abundance before immunoprecipitation to get the enrichment. To calculate relative enrichment, we divided enrichment of the mutated peptide by enrichment of the wild-type peptide. Since most of the single-mutant peptides had wild-type levels of enrichment, we averaged enrichment of the wild-type peptide enrichment with the middle two quartiles of enrichment of single-mutant peptides to get a better estimate of the wild-type peptide enrichment.

Respiratory syncytial virus and Herpesvirus 1 and 2 serology

Serum from 44 donors was tested for Herpesvirus 1 and Herpesvirus 2 antibodies using the HerpeSelect® 1 and 2 Immunoblot IgG kit (Focus Diagnostics) according to manufacturer's protocol. Serum from 60 donors was tested for Respiratory syncytial virus antibodies using Anti-Respiratory syncytial virus (RSV) IgG Human ELISA Kit (ab108765) according to manufacturer's protocol.

Chapter 4:

Identification of a novel subclass of scleroderma with autoantibodies against the minor spliceosome complex

Introduction

Scleroderma, also known as systemic sclerosis (SSc), is a chronic autoimmune rheumatic disease with a prevalence of 50 to 300 cases per 1 million persons and an incidence of 2.3 to 22.8 cases per 1 million persons per year (*97*). Patients with SSc develop a complex mixture of pathologies including vasculopathy and fibrosis of the skin and internal organs which can lead to death, most commonly by involvement of the heart and lungs leading to pulmonary fibrosis and pulmonary arterial hypertension (PAH) (*98*). The exact cause of SSc is still unknown, but the early stages of disease are believed to be triggered by endothelial cell injury and apoptosis (*24*).

SSc has been classified as an autoimmune disease because autoantibodies are frequently found in patients. In fact, approximately 90% of patients with SSc have antinuclear antibodies (*24*). Researchers have discovered that up to 50% of patients with SSc fall into one of three generally mutually exclusive subclasses which have autoantibodies targeting either the centromere, topoisomerase 1, or RNA polymerase III (Pol III) (*25*). Several reports have demonstrated different prognoses associated with these three subclasses (*97*, *99–101*), but the origin of the autoantibodies and their role, if any, in the initiation and progression of disease is still largely unknown.

The results of a recent study suggested that anti-Pol III autoantibodies may arise from a specific T cell response against a tumor mutation-associated neoantigen which primes cross-reactive antibody response against the wild-type version of Pol III (*26*). The study was motivated by a previous observation that patients with SSc and anti-Pol III autoantibodies tend to have coincident onset of SSc and cancer (*27*). The same was not true for patients with autoantibodies against either the centromere or topoisomerase1, but many patients with autoantibodies that did not fall into any of the three subclasses did have coincident cancer. We decided to study sera

from these patients to look for novel autoantibody specificities that may be associated with coincident cancer using the Phage-Immunoprecipiation Sequencing (PhIP-Seq) and Parallel Analysis of Translated Open Reading Frames (PLATO) techniques previously developed by our laboratory.

Both techniques can identify the antigen targets of antibodies in serum by immunoprecipitation in the presence of a mixture of potential protein antigens, each of which is associated with the DNA that encodes it. High-throughput sequencing of the mixture of DNA molecules before and immunoprecipitation reveals which protein antigens were enriched due to binding by antibodies in the sample. PhIP-Seq and PLATO differ in the method of protein display and the type of DNA encoding the displayed protein. In PhIP-Seq, the DNA are oligonucleotides originally synthesized on programmable microarrays that are displayed on the surface of bacteriophage as fusions to its coat protein (*19*). In PLATO, the DNA are a collection of full-length open reading frames that are expressed using ribosome display (*85*).

For PhIP-Seq, our laboratory has a collection of oligonucleotides encoding 90 amino acid protein fragments tiling through the entire human proteome with 45 amino acid overlap cloned into a T7 bacteriophage display vector. This reagent offers a complete and much more uniform representation of the human proteome compared to traditional cDNA expression libraries previously used to identify unknown autoantibody targets. However, because high-yield synthesis of longer DNA sequences at scale is not yet possible, PhIP-Seq is limited to display of protein fragments. Many antibodies recognize discontinuous epitopes that may not be captured in a 90 amino acid protein fragment. The PLATO technique, which displays full-length proteins,

overcomes this limitation. However, none of the human open reading frame collections contain the complete proteome, so there is gaps in PLATO's coverage.

We used these two complementary techniques to screen serum samples from patients with SSc and several others, including healthy donors and patients with other autoimmune diseases. We looked for novel autoantibody specificities in the patients with SSc who did not belong to any of the three known subclasses and determined if any newly identified subclasses were associated with coincident cancer.

Results

Confirmation of anti-Pol III autoantibody status

In our collaboration with Antony Rosen and Livia Casciola-Rosen at the Johns Hopkins School of Medicine, we obtained 48 serum samples from patients with SSc. Of these patients, 32 had autoantibodies against Pol III, as detected by prior serum ELISA or immunoprecipitation-Western blot (IP-WB) analysis. The remaining 16 did not have detectable autoantibodies against Pol III, topoisomerase 1, or centromeres but did have coincident cancer onset. We first assayed all of these samples using PhIP-Seq to identify their autoantibody specificities.

To assess PhIP-Seq's performance on these samples, we analyzed our ability to detect the known autoantibodies against Pol III. We detected autoantibodies against at least one subunit of Pol III in 27 of the 32 samples that had known anti-Pol III autoantibodies and 0 of the 16 samples that did not (Figure 30). These 32 samples were originally identified as positive for anti-Pol III autoantibodies using serum ELISA or IP-WB with the polymerase (RNA) III (DNA

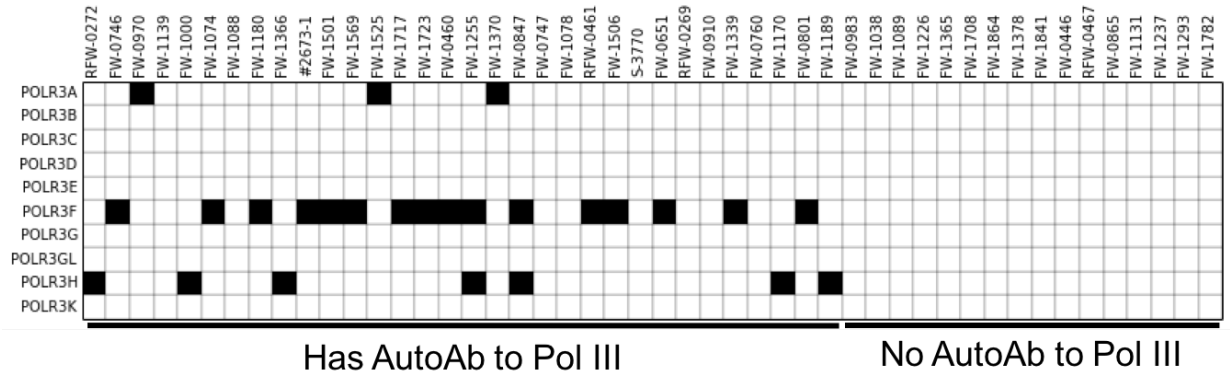

Figure 30. Detection of autoantibodies to Pol III subunits in sera from patients with SSc. Each column represents a patient with SSc that was screened using the PhIP-Seq assay. The labeled bars below the chart indicate whether or not the patient has detectable autoantibodies to Pol III. Each row is a gene that encodes a subunit of the Pol III complex. Each cell is colored black if serum from that column's patient contains autoantibodies to any protein fragment of that row's gene as detected by PhIP-Seq.

directed) polypeptide A (POLR3A) subunit of the Pol III complex. PhIP-Seq's performance on the POLR3A subunit alone was actually relatively poor. We detected autoantibodies against POLR3A in only 3 of the 32 samples. As briefly mentioned in this chapter's introduction, the discrepancy between the results of PhIP-Seq versus ELISA or IP-WB is likely because the bacterially expressed 90 amino acid protein fragments used in PhIP-Seq will not capture antibodies that recognize discontinuous epitopes or post-translationally modified epitopes. In contrast, ELISA and IP-WB assays using the full-length protein expressed in eukaryotic cells will be more sensitive for detecting these antibodies. We wanted to see if PLATO would be more sensitive than PhIP-Seq because it uses full-length proteins, but unfortunately POLR3A is not in our collection of human ORFs.

However, our overall performance for detecting anti-Pol III autoantibodies using PhIP-Seq was still high because we frequently detected autoantibodies to other Pol III subunits,

specifically POLR3F and POLR3H (Figure 30). Subsequent ELISA and IP-WB experiments with POLR3F and POLR3H confirmed the prevalence of autoantibodies against these antigens. These results are likely evidence of epitope spreading of the autoantibody response within the Pol III complex (*102*). This signature gives greater confidence that the observed autoantibodies are part of a true autoimmune response rather than a spurious result due to cross-reactive antibody binding or multiple hypothesis testing.

Identification of novel subclass with autoantibodies against the minor spliceosome complex

Satisfied that the PhIP-Seq assay has high performance on the samples, we next sought to identify novel autoantibody specificities in the set of samples from patients without autoantibodies against Pol III, topoisomerase 1, or centromeres. To accomplish this, we ranked all of the human proteins based on how frequently autoantibodies against them were observed in the 16 serum samples without known autoantibody specificities compared to the 32 samples with known autoantibody specificities. There were many candidate autoantigens that appeared to be more frequent in the set of 16 samples with unknown autoantigens (Figure 31).

In order to prioritize the candidates for follow-up analysis, we looked for evidence of intramolecular epitope spreading. Because PhIP-Seq is performed using protein fragments rather than full-length proteins, we are able to discern the presence of multiple antibodies that recognize distinct epitopes on the same protein. The candidate autoantigens in which we detect intramolecular epitope spreading are more likely to be the result of a true autoimmune response and thus should be prioritized for follow-up.

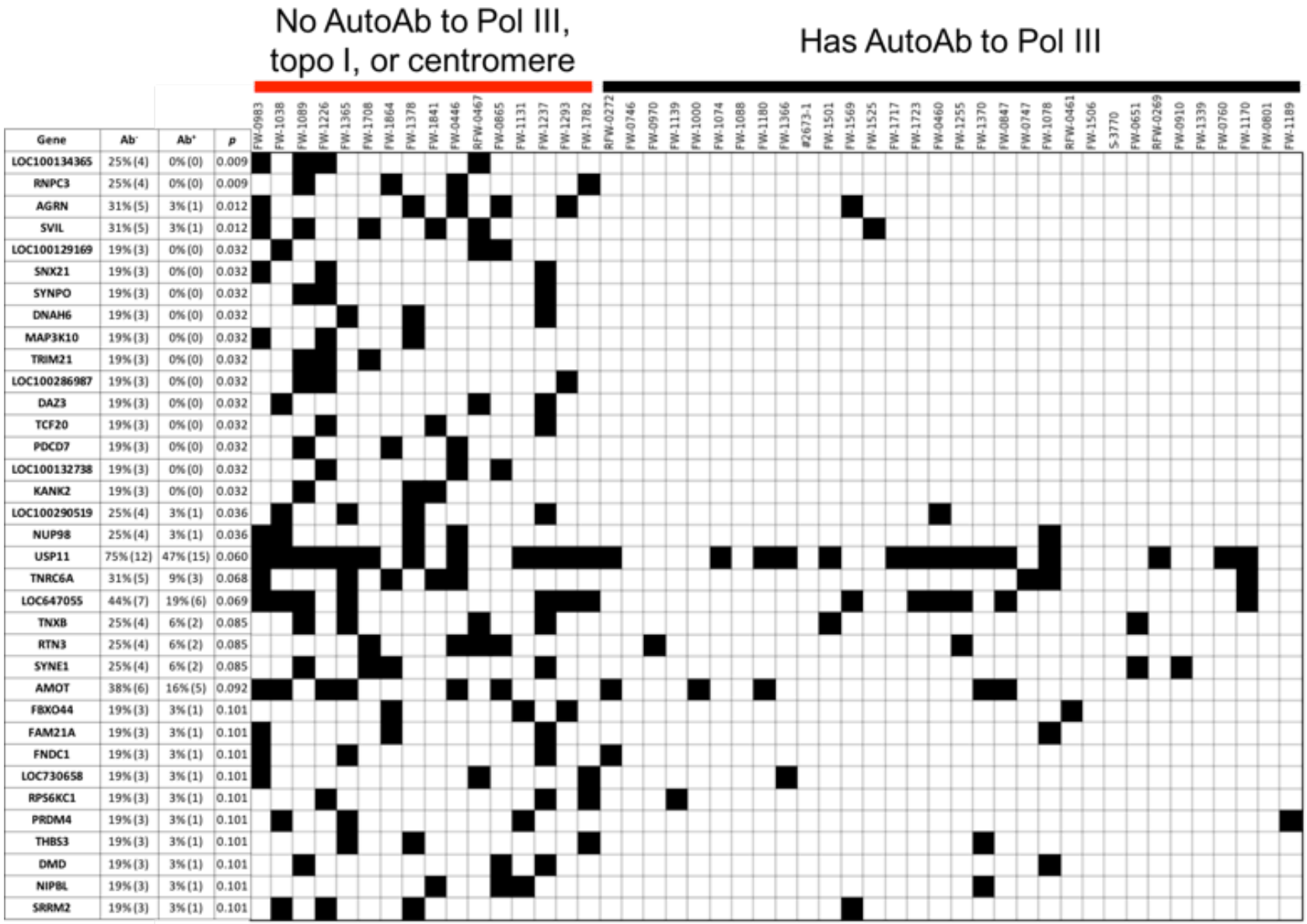
No AutoAb to Pol III, topo I, or centromere
Has AutoAb to Pol III
Gene
Ab⁻
Ab⁺
p
LOC100134365 25% (4) 0% (0) 0.009
RNPC3 25% (4) 0% (0) 0.009
AGRN 31% (5) 3% (1) 0.012
SVIL 31% (5) 3% (1) 0.012
LOC100129169 19% (3) 0% (0) 0.032
SNX21 19% (3) 0% (0) 0.032
SYNPO 19% (3) 0% (0) 0.032
DNAH6 19% (3) 0% (0) 0.032
MAP3K10 19% (3) 0% (0) 0.032
TRIM21 19% (3) 0% (0) 0.032
LOC100286987 19% (3) 0% (0) 0.032
DAZ3 19% (3) 0% (0) 0.032
TCF20 19% (3) 0% (0) 0.032
PDCD7 19% (3) 0% (0) 0.032
LOC100132738 19% (3) 0% (0) 0.032
KANK2 19% (3) 0% (0) 0.032
LOC100290519 25% (4) 3% (1) 0.036
NUP98 25% (4) 3% (1) 0.036
USP11 75% (12) 47% (15) 0.060
TNRC6A 31% (5) 9% (3) 0.068
LOC647055 44% (7) 19% (6) 0.069
TNXB 25% (4) 6% (2) 0.085
RTN3 25% (4) 6% (2) 0.085
SYNE1 25% (4) 6% (2) 0.085
AMOT 38% (6) 16% (5) 0.092
FBXO44 19% (3) 3% (1) 0.101
FAM21A 19% (3) 3% (1) 0.101
FNDC1 19% (3) 3% (1) 0.101
LOC730658 19% (3) 3% (1) 0.101
RPS6KC1 19% (3) 3% (1) 0.101
PRDM4 19% (3) 3% (1) 0.101
THBS3 19% (3) 3% (1) 0.101
DMD 19% (3) 3% (1) 0.101
NIPBL 19% (3) 3% (1) 0.101
SRRM2 19% (3) 3% (1) 0.101
FW-0983
FW-1038
FW-1089
FW-1226
FW-1365
FW-1708
FW-1864
FW-1378
FW-1841
FW-0446
RFW-0467
FW-0865
FW-1131
FW-1237
FW-1293
FW-1782
RFW-0272
FW-0746
FW-0970
FW-1139
FW-1000
FW-1074
FW-1088
FW-1180
FW-1366
#2673-1
FW-1501
FW-1569
FW-1525
FW-1717
FW-1723
FW-0460
FW-1255
FW-1370
FW-0847
FW-0747
FW-1078
RFW-0461
FW-1506
S-3770
FW-0651
RFW-0269
FW-0910
FW-1339
FW-0760
FW-1170
FW-0801
FW-1189

Figure 31. Ranked list of the top candidate novel autoantigens in SSc. Each column of the chart represents a patient with SSc that was screened using the PhIP-Seq assay. The labeled bars above the chart indicate whether or not the patient has detectable autoantibodies to Pol III. Each row of both the chart and the adjacent table represents a patient serum sample. Each cell is colored black if serum from that column's patient contains autoantibodies to any protein fragment of that row's gene as detected by PhIP-Seq. The column "Ab^-" in the table lists the percentage of samples without known autoantibody specificities that have autoantibodies against that row's gene product. The column "Ab^+" in the table lists the same except for the samples with known autoantibody specificities. Finally, the column "*p*" in the table lists the p-value of a Fisher's exact test for whether the autoantibodies against row's gene product is more frequent in the samples without known autoantibody specificities.

In performing the described analyses, we found that the strongest candidate autoantigen on our list was RNPC3. We detected auto-antibodies against RNPC3 in 4 of the 16 samples without known autoantibody specificities and none of the 32 samples with known specificities. Including our results from other screens using PhIP-Seq, we found that autoantibodies against RNPC3 were not present in any of 123 serum samples, including ones from healthy donors and patients with the autoimmune diseases IgG4-related disease and dermatomyositis. So autoantibodies against RNPC3 appear to be a specific marker of SSc. In addition, we saw evidence of epitope spreading within RNPC3 in all four of the samples with autoantibodies against RNPC3 (**Error! Reference source not found.**). We observed autoantibodies against the same four consecutive protein fragments. This region is not repetitive; the four fragments do not have any stretch of four or more amino acids in common. Because the tiling protein fragments are 90 amino acids long, with 45 amino acid overlap, detection of autoantibodies against four consecutive fragments indicates there are at least two distinct antibodies recognizing separate epitopes. Thus, the observed pattern of autoantibodies against RNPC3 is likely the result of

intramolecular epitope spreading, suggesting there is a true autoimmune response against RNPC3.

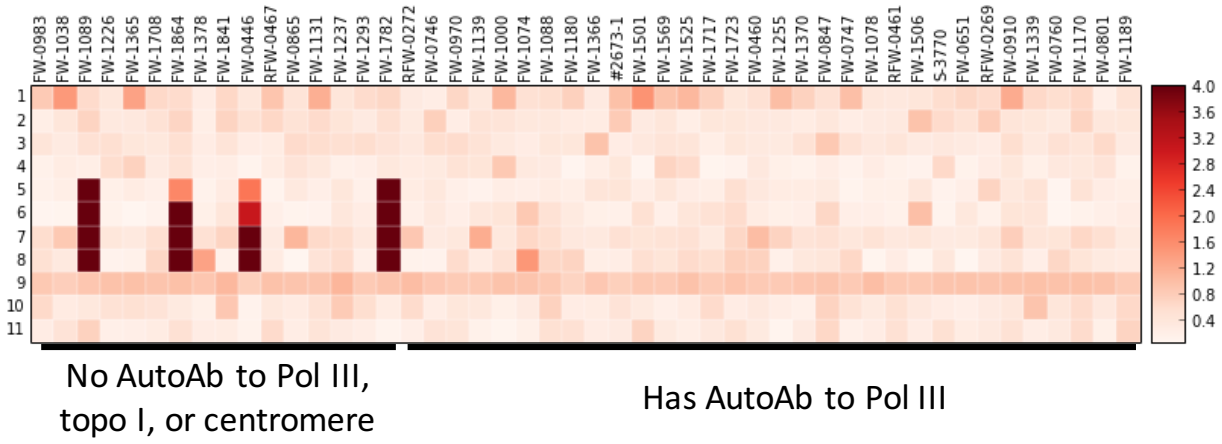

Figure 32. Evidence for intramolecular epitope spreading in RNPC3. Each column of the chart represents a patient with SSc that was screened using the PhIP-Seq assay. The labeled bars before the chart indicate whether or not the patient has detectable autoantibodies to Pol III. Each row of the chart represents one of the 90 amino acid protein fragments tiled through the entire RNPC3 gene (1 is the N-terminal fragment). The color of each cell represents the $-\log_{10}$(p-value) for enrichment of that row's protein fragment in that column's sample. Greater confidence in the detection of an autoantibody is indicated by a darker color (larger $-\log_{10}$(p-value)), as labeled on the scale in the colorbar to the right of the chart.

RNPC3 encodes RNA-binding protein 40, a member of the minor spliceosome complex which participates in removal of U12-type introns from pre-mRNA (*103*, *104*). In three of the four samples with autoantibodies against RNPC3, we also detected autoantibodies against PDCD7, another member of the minor spliceosome complex (Figure 31). This data suggests that in addition to intramolecular epitope spreading within RNPC3, we detected intermolecular epitope spreading within the minor spliceosome complex, providing further evidence that we observed a true autoimmune response.

Confident that we were observing a true autoantibody response against RNPC3 that was specific to patients with SSc, we asked our collaborators, Antony Rosen and Livia Casciola-Rosen, to perform IP-WBs using RNPC3 with these and other samples from patients with SSc.

	Coincident Cancer	No coincident cancer	Totals
Has autoantibodies to Pol III, topo 1, or centromere	0/25	0/42	0/67
No autoantibodies to Pol III, topo 1, or centromere	10/69	8/61	18/130
Totals	10/94	8/103	18/197

Table 6. Patients with autoantibodies against RNPC3 are a novel subclass of SSc. The numbers in the table represent the fraction of samples with autoantibodies to RNPC3, with the denominator representing the total number of samples that fit the criteria specified in the row and column labels. The main rows split the samples into ones that have known autoantibody specificities and ones that do not. The main columns split the samples into ones that have coincident cancer and ones that do not. The marginal rows and columns are totals summed across the rows or columns, respectively.

Their results confirmed our findings in this set of 48 samples and also found 14 additional cases of patients with SSc with autoantibodies against RNPC3 among other samples in their repository. In total, 18 of 197 patients with SSc that were tested by IP-WB had autoantibodies against RNPC3 (**Error! Reference source not found.**). Similar to what has been observed with the three known subclasses, this novel subclass was mutually exclusive with all of the other subclasses. None of the 18 had autoantibodies against Pol III, topoisomerase 1, or centromeres. We also looked to see if autoantibodies against RNPC3 were associated with coincident cancer, as had previously been observed for patients with autoantibodies against POLR3A (*27*). There was a very slightly higher frequency of coincident cancer in patients with autoantibodies against (10 with coincident cancer versus 8 without coincident cancer), but the

difference was not statistically significant (**Error! Reference source not found.**). However, these data do not preclude the possibility that the 8 patients did indeed have coincident malignancies, but that the nascent tumors were controlled by the immune system before they progressed enough to be diagnosed.

		RNPC3+		not RNPC3+		RNPC3+ vs not RNPC3+
		+	-	+	-	Fisher p-value
gene	position					
RNPC3	8	15	3	0	123	1.333543e-17
SNRNP48	2	4	14	0	123	1.939508e-04
RNPC3	6	4	14	0	123	1.939508e-04
OR52N1	2	4	14	4	119	9.640935e-03
RNPC3	7	2	16	0	123	1.550152e-02
DDX19B	1	2	16	0	123	1.550152e-02
	2	2	16	0	123	1.550152e-02
TOX	2	2	16	0	123	1.550152e-02
DDX19A	1	2	16	0	123	1.550152e-02
SRCIN1	21	2	16	0	123	1.550152e-02
KDM2A	10	2	16	0	123	1.550152e-02
TCF3	4	2	16	0	123	1.550152e-02
REPS2	4	9	9	28	95	1.855276e-02
ZFHX4	23	2	16	1	122	4.293586e-02
LOC727828	6	2	16	1	122	4.293586e-02
TIAL1	1	2	16	1	122	4.293586e-02
PTGIR	6	2	16	1	122	4.293586e-02
CD2BP2	2	10	8	42	81	6.901847e-02
LOC100134244	7	5	13	14	109	7.012544e-02
LOC100132351	3	2	16	2	121	7.931619e-02

Table 7. Top candidate autoantigens associated with the subclass of patients with SSc who have autoantibodies against RNPC3 detected by PhIP-Seq. The first two columns indicate the gene and position of the protein fragment each row refers to. The first position is the N-terminal 90 amino acid fragment and each consecutive position is the next 90 amino acid fragment with 45 amino acid overlap. The "+" and "-" columns under the "RNPC3+" heading indicate the number of samples with autoantibodies against RNPC3 that did or did not have detectable autoantibodies against that row's protein fragment, respectively. The "+" and "-" columns under the "not RNPC3+" heading represent the same except for the 123 control samples. The final column is the p-value for a Fisher's exact test for whether autoantibodies against that row's protein fragment is more frequent in the "RNPC3+" group compared to the "not RNPC3+" group.

	Fold-Change			
Gene	FW-1089	FW-0446	FW-1782	FW-1864
RNPC3	22.6	10.5	8.0	1.5
SNRNP25	9.2	3.5	9.6	0.9
ARHGAP27	6.4	1.3	1.1	1.1
SNRNP35	5.4	3.7	4.5	1.4
SEPT9	4.5	0.9	1.1	0.8
TMEM175	4.3	1.6	2.8	1.6
SH3D19	4.3	0.9	1.1	1.4
SPATA24	4.3	0.7	0.8	0.7
SCG3	4.1	0.9	1.5	0.3
ANXA7	4.0	1.1	0.8	2.1
ACVR2A	3.7	1.2	1.0	1.8
APOF	3.7	0.9	0.9	1.6
ACVR2A	3.6	1.2	1.1	1.9
C1orf125	3.6	1.4	1.8	2.0
MFSD9	3.6	3.0	2.6	4.6
PLCB1	3.5	1.1	1.1	1.2
POR	3.4	1.4	2.0	2.7
SR140	3.4	1.4	1.2	1.0
ANKRD12	3.3	0.8	0.7	0.4
CREB3L4	3.3	2.4	1.0	3.3
FAM133A	3.3	2.1	2.3	0.8
ADCYAP1R1	3.3	1.0	1.0	1.1
PCSK4	3.3	0.4	0.4	1.4

Table 8. Top candidate autoantigens detected by PLATO in the four patients with SSc who have autoantibodies against RNPC3. Each row shows data corresponding to the candidate autoantigen gene listed in the first column. The remaining four columns show the fold-change in relative abundance for each gene after immunoprecipitation with patient serum compared to beads alone (see Methods). The data in each of the columns were obtained from the sample identified in the column label. The data are sorted in descending order based on fold-change in sample "FW-1089", but RNPC3, SNRNP25, and SNRNP35 are in the top 5 greatest fold-change in "FW-0446" and "FW-1782".

We performed two more screens to determine if there are any more autoantibody specificities associated with this novel RNPC3 subclass. For the first screen, we used PhIP-Seq and looked for autoantibody specificities that were more frequent in the 18 samples with autoantibodies against RNPC3, as identified by IP-WB analysis using recombinant RNPC3

antigen, than in 123 control samples, including samples from health donors and patients with other autoimmune diseases such as IgG4-related disease and dermatomyositis. These 18 samples included the four in which PhIP-Seq had previously detected autoantibodies against three to four peptides each, at least two of which were non-overlapping (Figure 32). As expected, three of the top five candidates were peptides derived from RNPC3 (**Error! Reference source not found.**). These autoantigenic peptides were the same ones detected in the original four samples, and several of the samples had autoantibodies against multiple of these peptides. The top non-RNPC3 candidate was a peptide from SNRNP48, which, like RNPC3 and PDCD7 previously, is a component of the minor spliceosome complex.

For the second screen, we used PLATO on the four samples we had originally identified with autoantibodies against RNPC3. In three of the four samples, RNPC3 was the most or second most strongly enriched ORF, further confirming our PhIP-Seq results (**Error! Reference source not found.**). The remaining sample had poor enrichment across all of the ORFs, perhaps because there was an issue with RNA degradation in that reaction. In addition to RNPC3, we found that SNRNP25 and SNRNP35, both components of the minor spliceosome complex, were also significantly enriched in the same three samples. Our PhIP-Seq screens did not detect significant enrichment of protein fragments derived from these two proteins in these three samples, perhaps because of PLATO is more sensitive for detecting antibodies against discontinuous epitopes, as discussed in this chapter's introduction.

Combining the results of all our screens, we discovered that patients in this novel subclass of SSc have autoantibodies against many components of the minor spliceosome complex (Table 9). These results provide strong evidence that this novel subclass of patients with

SSc share the distinct pathological feature of an autoimmune humoral response against the minor spliceosome complex.

Protein (Gene)	Autoantibodies detected by PhIP-Seq	Autoantibodies detected by PLATO
Sm proteins[1]		
SF3b complex[2]		
20K (ZMAT5)		
25K (SNRNP25)		✓
31K (ZCRB1)		
35K (SNRNP35)		✓
48K (SNRNP48)	✓	
59K (PDCD7)	✓	
65K (RNPC3)	✓	✓
Urp (ZRSR2)		
hPrp43 (DHX15)		
Y Box-1 (YBX1)		

Table 9. Discovery of autoantibodies against multiple components of the minor spliceosome complex. Each row represents one of the protein components of the minor spliceosome complex (*105*) as indicated by the label in the first column (gene names in parentheses). Checkmarks in next two columns indicate whether autoantibodies to that protein were identified by PhIP-Seq, PLATO, or both.

[1] Sm proteins B/B', D1, D2, D3, E, F, and G

[2] multi-subunit complex

Discussion

In this study, we have demonstrated the power of combining the two complementary approaches developed in our laboratory, PhIP-Seq and PLATO, for identifying autoantibody specificities in large, heterogeneous populations. Although PhIP-Seq has the advantages of being high throughput and a complete synthetic representation of the human proteome at the protein fragment level, it can miss antibodies that recognize discontinuous epitopes. In contrast, PLATO is lower throughput and does not provide complete coverage of the human proteome in its

current form, but will detect antibodies against discontinuous epitopes. Thus, although both techniques could detect autoantibodies against RNPC3, other autoantibody specificities were only detectable by either only PhIP-Seq or only PLATO.

Combining the results from the PhIP-Seq and PLATO screens revealed a novel subclass of patients with SSc with autoantibodies against the minor spliceosome complex. Our results indicate significant intramolecular epitope spreading within RNPC3 and intermolecular epitope spreading within the entire complex. This immunological signature is highly indicative of a true autoimmune response and allowed us to prioritize the RNPC3 and the minor spliceosome for follow-up analysis by IP-WB using full-length antigens that confirmed our initial findings. These results illustrate the usage of immunological understanding to recognize candidate autoantigens that are unlikely to simply be due to cross-reactivity or false positives.

One striking feature of our results is that when we used PhIP-Seq to map the autoantibody epitopes in the 18 samples with autoantibodies against RNPC3 as detected by Western Blot, we found that almost all of these samples had autoantibodies against the same set of peptides in RNPC3. This pattern is highly reminiscent of our results with Virscan, in which the vast majority of people generated antibodies against the same viral peptides following exposure. As discussed in the preceding chapter, it is possible that this region of the protein exhibits particularly antigenic properties. Although not repetitive, the targeted region of RNPC3 is fairly rich in amino acids that we found to be enriched in viral epitope determinants (12.5% P, 12% E, 8.5% K, 7.1% D; Figure 17). It is possible that the antigenic features of this region of the protein elicit an immunodominant antibody response across individuals. However, our results do

not rule out the possibility that highly similar antibodies arose in all of these individuals from a shared naïve B cell precursor, as discussed in the preceding chapter.

The initial trigger for SSc is believed to be damage to the endothelial cells. Because the minor spliceosome complex is a nuclear antigen and should not be exposed on undamaged endothelial cells, autoantigens against the complex may not be directly involved in the initiation of disease. However, upon initiation of endothelial cell death, the nuclear antigens are released into the extracellular environment and could elicit an autoimmune response. Unfortunately, our data does not reveal whether the antinuclear autoantibodies are involved in pathogenesis or are merely a consequence of disease, a question which is still unanswered for most rheumatic diseases (*106*). Unlike patients with systemic lupus erythematosus, patients with SSc do not experience deposition of highly inflammatory immune complexes containing autoantibodies in affected tissues (*107*). However, treating patients with SSc with the B cell depleting monoclonal antibody Rituximab improves skin fibrosis and prevents worsening lung fibrosis, suggesting that the humoral response does play a role in pathogenesis (*108*).

In the case of patients with SSc in the subclass with autoantibodies against Pol III, the autoantibodies appear to be a cross-reactive B cell response generated after a specific T cell response to a mutant Pol III neoantigen in a tumor (*109*). This model suggests that in these patients SSc develops due to aberrant tumor immunity and autoantibodies against Pol III play an early role in pathogenesis. It is possible a similar mechanism operates in the subclass with autoantibodies against the minor spliceosome complex. Although only half of the patients in this subclass had coincident cancer diagnosis, it is possible that the other half's autoimmunity was also triggered by a tumor, but the tumor regressed before being diagnosed. Sequencing the genes

encoding members of the minor spliceosome in the tumors of patients in this subclass could also reveal tumor neoantigen driven autoimmunity.

One interesting observation is that all of the subclasses of patients with SSc are mutually exclusive. It is possible this result is due to heterogeneity in the underlying cause. Perhaps SSc in the POLR3A subclass are all initiated by immunity against tumors with POLR3A mutations, whereas SSc in other subclasses are initiated by other mechanisms that elicit different autoimmune specificities. It is also possible that there is a genetic component. Patients with particular HLA alleles or polymorphisms in other immune-related genes may be predisposed to generate particular autoantibody specificites. Genome-wide associations studies have found that particular SNPs in HLA class II genes are associated with autoantibodies against topoisomerase 1 or centromeres (*110*).

Regardless of the mechanism of their origin, autoantibodies against the minor spliceosome complex could serve as a useful diagnostic and prognostic biomarker. It is already known that patients in the three known subclasses of SSc have very different clinical prognoses (*97*, *99–101*). We can now study the medical records or conduct longitudinal studies of patients in the novel subclass to identify distinguishing clinical features. Knowledge of these clinical features may also inform understanding of the underlying pathogenesis.

In this work, we used a pair of complementary antigen discovery technologies developed in our laboratory and discovered a novel subclass of patients with SSc who have autoantibodies against the minor spliceosome complex. These autoantibodies can serve as diagnostic and prognostic biomarkers for clinical care. Because over half of the patients in this subclass had coincident cancer diagnoses, we believe the autoantibodies may result from an anti-tumor

neoantigen immune response that cross reacts with the wild type, as previously reported for a known subclass of patients with SSc.

Methods

PhIP-Seq: library design, construction, assay, and analysis

The PhIP-Seq autoantigen library sequences were derived from the RefSeq collection of human protein sequences and cloned into a T7 bacteriophage display system similar to what has been previously described (*19*). Briefly, the protein sequences for 90 amino acid protein fragments tiling with 45 amino acid overlap through all of the human proteins in RefSeq were synthesized as 300 bp oligonucleotides with 5' and 3' adapter sequences using a programmable DNA microarray. These oligonucleotides were then cloned into the T7-Select 10-3b bacteriophage display vector and amplified in the BLT5403 according to the manufacturer's protocol (Merck Millipore). Aliquots are frozen at -80ºC in 10% DMSO until use.

PhIP-Seq assays were conducted as previously described (*19*). First, we blocked each well of a 96 deep-well plate with 1 mL of 3% BSA in TBST overnight on a rotator at 4ºC. To each pre-blocked well, we added sera or plasma containing approximately 2 μg of IgG (quantified using a Human IgG ELISA Quantitation Set (Bethyl Laboratories)) and 1 mL of the bacteriophage library diluted to approximately 2×10^5 fold representation (2×10^{10} pfu for a library of 10^5 clones) in phage extraction buffer (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 6 mM $MgSO_4$). We performed two technical replicates for each sample. We allowed the antibodies to bind the phage overnight on a rotator at 4ºC. The next day, we added 20 μL each of magnetic Protein A and Protein G Dynabeads (Invitrogen) to each well and allowed immunoprecipitation to occur for 4 h on a rotator at 4ºC. Using a 96-well magnetic stand, we then washed the beads

three times with 400 µL of PhIP-Seq wash buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.1% NP-40). After the final wash, we resuspended the beads in 40 µL of water and lysed the phage at 95 ºC for 10 m. We also lysed phage from the library before immunoprecipitation ("input") and after immunoprecipitation with beads alone.

We prepared the DNA for multiplexed Illumina sequencing using a slightly modified version of a previously published protocol (*96*). We performed two rounds of PCR amplification on the lysed phage material using hot start Q5 polymerase according to the manufacturer suggested protocol (NEB). The first round of PCR used the primers IS7_HsORF5_2 (ACACTCTTTCCCTACACGACTCCAGTCAGGTGTGATGCTC) and IS8_HsORF3_2 (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCGAGCTTATCGTCGTCATCC). The second round of PCR used 1 µL of the first round product and the primers IS4_HsORF5_2 (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACTCCAGT) and a different unique indexing primer for each sample to be multiplexed for sequencing (CAAGCAGAAGACGGCATACGAGATxxxxxxxGTGACTGGAGTTCAGACGTGT, where "xxxxxxx" denotes a unique 7 nt indexing sequence). After the second round of PCR, we determined the DNA concentration of each sample by qPCR and pooled equimolar amounts of all samples for gel extraction. Following gel extraction, the pooled DNA was sequenced by the Harvard Medical School Biopolymers Facility using a 50 bp read cycle on an Illumina HiSeq 2000 or 2500. We pooled up to 96 samples for sequencing on each lane and generally obtained approximately 100 - 200 million reads per lane (1,000,000 to 2,000,000 reads per sample).

We performed the initial informatics and statistical analysis using a slightly modified version of the previously published technique (*19*, *70*). We first mapped the sequencing reads to

the original library sequences using Bowtie and counted the frequency of each clone in the "input" and each sample "output" (*62*). Since the majority of clones are not enriched we use the observed distribution of output counts as a null distribution. We found that a zero-inflated generalized poisson distribution fits our output counts well. We use this null distribution to calculate a p-value for the likelihood of enrichment for each clone. The probability mass function for the zero-inflated generalized poisson distribution is

$$P(Y = y) = \begin{cases} \pi + (1-\pi)(\theta(\theta+\lambda)^{x-1}e^{-\theta-x\lambda}), & y = 0 \\ (1-\pi)(\theta(\theta+\lambda)^{x-1}e^{-\theta-x\lambda}), & y > 0 \end{cases}$$

We used maximum likelihood estimation to regress the parameters π, θ, and λ to fit the distribution of counts after immunoprecipitation for all clones present at a particular frequency count in the input. We repeated this procedure for all of the observed input counts and found that θ and λ are well fit by linear regression and π by an exponential regression as a function of input count (Figure 9). Finally, for each clone we used its input count and the regression results to determine the null distribution based on the zero-inflated generalized poisson model, which we used to calculate the -$\log_{10}$(p-value) of obtaining the observed count.

To call hits, we determined the threshold for reproducibility between technical replicates based on a previously published method (*70*). Briefly, we made scatter plots of the log10 of the -log10 (p-values) and used a sliding window of width 0.005 from 0 to 2 across the axis of one replicate. For all the clones that fell within each window, we calculated the median and median absolute deviation of the log10 of the –log10 (p-values) in the other replicate and plotted it against the window location (Figure 10). We called the threshold for reproducibility the first window in which the median was greater than the median absolute deviation. We found that the

distribution of the threshold –log10 (p-value) was centered around a mean of approximately 2.3. So we called a peptide a hit if the –log10 (p-value) was at least 2.3 in both replicates.

PLATO: library construction, assay, and analysis

The PLATO library is a barcoded version of the one previously described (*85*). The vector pRDDEST described in that paper was modified by cloning a stretch of 30 random nucleotides (N's) following the attB2 site and in-frame with the downstream TolA sequence. These barcodes were previously selected to contain only sequences without a stop codon by cloning them, in-frame, as 5' fusions to an antibiotic resistance gene, transforming them into sensitive bacteria, and selecting for resistance. The human ORFeome v5.1 collection was cloned into this barcoded vector using Gateway Cloning as previously described (*111*). The resulting DNA was electroporated into DH10B cells according the manufacturer's protocol (Invitrogen), which were grown on carbenicillin containing LB agar plates, and plasmid DNA was maxi-prepped from the cells scraped from these plates and stored at -20ºC until use.

The PLATO assays were performed as previously described (*111*). Briefly, the plasmid DNA was PCR amplified using the T7B (5'-ATACGAAATTAATACGACTCACTATAGGGA GACCACAACGG-3') and TolAK (5'-CCGCACACCAGTAAGGTGTGCGGTTTCAGTTGC CGCTTTCTTTCT-3') primers. The amplified DNA is PCR purified and *in vitro* transcribed using the RiboMAX large-scale RNA production system-T7 kit according to the manufacturer's protocols (Promega). The RNA is purified using MegaClear according to the manufacturer's protocols (Ambion) and 15 µg is used for 100 µL *in vitro* translation reaction using the RTS 100 *E. coli* HY kit according the manufacturer's protocols (5 Prime). 12.5 µL of the *in vitro* translation reaction is diluted in 85.5 µL of Selection Buffer [2.5 mg/mL heparin, 1% (wt/vol) BSA, and 83.3 µg/mL yeast tRNA in 50 mM Tris acetate and 150 mM NaCl, pH 7.5, DEPC

treated] and added to 40 µL of a 1:1 mix of Protein A and Protein G Dynabeads (Life Technologies) that have bound 2 µg of IgG from serum overnight at 4ºC in PBST containing 1% (wt/vol) BSA and subsequently blocked in RD Selection buffer for 1 hour at room temperature. After 4 hours of incubation with the *in vitro* translation product at 4ºC, the beads are washed six times with 500 µL RD wash buffer (50 mM Tris acetate and 150 mM NaCl, pH 7.5, DEPC treated) and the remaining bound RNA is eluted with 50 µL EB20 (50 mM Tris acetate, 150 mM NaCl, 20 mM EDTA, pH 7.5) at 37ºC for 10 min. The eluted RNA is purified using the MegaClear kit according to the manufacturer's protocols (Ambion) and reverse transcribed with the TolART primer (5'-CGCTGCTTCTTCCGCAGCTTTAGC-3') using the SuperScript III kit according to the manufacturer's protocols (Life Technologies). The barcode region of the cDNA is then PCR amplified using the primers Adap-BCfor (5'-GTGACTGGAGTTCAGACGTGTGC TCTTCCGATCTACAAGTCACGTCCACAGTCGT-3') and P5-BCrev (5'-AATGATACGGC GACCACCGAACTACGGTGCGGCGAATATAC-3'). The second round of PCR used 1 µL of the first round product and the primers P5-BCrev and a different unique indexing primer for each sample to be multiplexed for sequencing (CAAGCAGAAGACGGCATACGAGATxxxxxxxGT GACTGGAGTTCAGACGTGT, where "xxxxxxx" denotes a unique 7 nt indexing sequence). After the second round of PCR, we determined the DNA concentration of each sample by qPCR and pooled equimolar amounts of all samples for gel extraction. Following gel extraction, the pooled DNA was sequenced by the Harvard Medical School Biopolymers Facility using a 50 bp read cycle on an Illumina HiSeq 2000 or 2500.

We first mapped the sequencing reads to the original library sequences using Bowtie and counted the frequency of each clone in the "input" and each sample "output" (*62*). We then used then used these sequencing read counts to calculate "fractional abundance" estimates for each

clone using the number of reads for that peptides divided by the total number of reads for that sample. The ratio of the fractional abundance in the output over the input is used as an estimate for fold-change enrichment.

Chapter 5:
Conclusion and Future Directions

Summary

In this chapter, I will summarize the results of the work contained in this thesis and suggest directions for future research.

In Chapter 2, I described a synthetic antibody library designed for high-throughput sequencing assisted selection. This method enables rapid *in vitro* selection of antibodies that bind specifically to a target of interest by bypassing the need for laborious single-clone screening for specific binding. We demonstrated application of this method to identify an antibody that binds specifically to a cancer-associated antigen.

In Chapter 3, I describe VirScan, a high-throughput assay for detection of antibodies against all known human viruses. We showed that VirScan has very high sensitivity and specificity for detecting a range of viruses and then applied it to over 500 serum samples from donors across four different continents. We found that the antiviral antibody response in different individuals targets strikingly similar epitopes, suggesting that there may be antibodies, or at least antibody specificities, which are shared across the global human population.

And last, in Chapter 4, I describe the use of the complementary approaches, PhIP-Seq and PLATO, to identify a novel subclass of patients with scleroderma. This subclass is mutually exclusive from the three previously known major subclasses and is characterized by autoantibodies against multiple components of the minor spliceosome complex. We also found some evidence that a significant portion of patients in this subclass may have developed autoimmunity in association with cancer.

Improving sequencing-assisted selection of affinity reagents

Although we were able to use our sequencing-assisted selection method to identify a synthetic antibody that specifically binds to the target of interest, the affinity of the antibody was rather weak. In addition, we found that the antibodies from our library were difficult to express at high yields. These drawbacks significantly limit the usefulness of our method. In this section, I describe ways to potentially overcome these limitations.

There are several methods to carry out *in vitro* affinity maturation of antibodies (*112*). Since our system uses ribosome display, one simple method would be to use an error-prone polymerase to amplify the cDNA after reverse transcription of the eluted RNA. The error-prone polymerase introduces random mutations into the population after each round of affinity purification. Higher affinity antibodies arise from the population of mutants of initially weakly binding antibody.

However, such random mutations pose an issue for high-throughput sequencing identification of individual antibodies. Because current high-throughput DNA sequencing technologies' read lengths are still too short to sequence an entire antibody gene, our sequencing-assisted method only reads the complementarity determining regions (CDRs). This is sufficient if the remaining portion of the antibody gene (framework sequences) is constant, which is true for our library design. However, it is not sufficient to identify antibodies if they contain important mutations introduced by an error-prone polymerase in the framework sequence, which is not sequenced.

One possible method to restrict mutations to the CDRs (which will likely confer the greatest improvements in antibody affinity) is to amplify the CDRs independently with error prone polymerase, then use a combination of splicing by overlap extension (*113*) and Gibson

assembly (*55*) (or traditional cloning) to assemble the CDRs into the original framework. However, after independent amplification and re-assembly, an antibody may contain CDRs from multiple different parent antibodies instead of just mutated CDRs of the single parent antibody. Shuffling of heavy- and light-chain pairs is a commonly used technique for affinity maturation, but shuffling CDRs within heavy chains may be detrimental.

Instead of affinity maturation during selection, another strategy is to improve the quality of the initial collection of synthetic antibodies. In particular, it is possible that only a relatively small proportion of the initial collection actually folded into functional antibodies. Previous structural studies showed that for a particular antibody framework, certain residues in the CDR-H2 loop are buried and may be necessary for proper folding (*114*). Our library also used a single framework, but the CDR-H2 loops were designed using a Hidden Markov Model based on crystal structures of antibodies of multiple different frameworks. In addition, because this implementation of Markov Models was inherently memory-less, it did not take into account any amino acid preference at a specific position. Thus, the CDR-H2 loops may not contain the proper residue at the normally buried positions for the specific framework we used, preventing the antibody from folding properly. If they constitute a significant percentage of the library, the effective size of the library would be much smaller and the probability of finding a high affinity antibody would be lower. It may be beneficial to design a library where only the solvent exposed residues on CDR-H2 are diversified.

Going beyond just antibodies, there are also alternative protein scaffolds for affinity reagents with much more favorable properties, including adnectins, affibodies, anticalins, and designed ankyrin repeats (*115*). It is possible to develop affinity reagents using these alternative

scaffolds which have very high affinity and specificity, but are much smaller than traditional antibodies (allowing full-length reads even with high-throughput sequencing), are more easily expressed at high yields even in bacteria, and do not require disulfide bonds and are thus functional intracellularly. Using these alternative scaffolds could improve the functional properties of the initial antibody library.

Understanding and exploiting recurrent viral B cell epitopes

One of the most striking findings in our VirScan studies was that almost everyone generates antibodies that recognize the same or highly similar epitopes after exposure to a virus. We found this to be true for several different populations and several different viruses. Our results did not elucidate a biological mechanism for this phenomenon, but we think that there are likely two possibilities.

The first possibility is that these recurrently targeted epitopes possess special properties that make them particularly antigenic. In our studies, we showed that peptides recognized by antibodies tended to share a slight compositional bias. However, we did not observe any statistically significant enrichment for surface antigens or terminal peptides among viral peptides recognized by antibodies. These results suggested that these properties of the peptide are not sufficient to explain highly recurrent epitopes, but it is a possibility that a combination of these properties perhaps in conjunction with properties we did not test for are critical determinants of antigenicity.

The second possibility is that the recurrently targeted epitopes we observe are actually the result of the same or highly similar antibodies arising across individuals. In fact, there is already some evidence for this possibility from high-throughput sequencing of antibody genes. Although

these methods rarely identify identical antibody sequences in unrelated individuals, even high-throughput sequencing only samples a small subset of the entire antibody repertoire so the absence of identical sequences may be due to undersampling (*7*). One study, which focused on analyzing very deep sequencing of antibodies against dengue virus, found that highly similar or even identical CDR-H3 sequences were present in multiple individuals (*95*). CDR-H3 is the most diverse CDR and the probability of observing such similarity by chance alone is infinitesimally small. Thus, these results suggest that the process for generating antibodies is not random but highly biased such that identical or nearly identical antibodies can frequently arise in multiple individuals. In fact, it is known that VDJ recombination and even addition of N and P nucleotides are highly biased such that certain rearrangements are much more highly favored (*94*).

Distinguishing between these possible explanations for our observations will require sequencing of the genes of antibodies from multiple individuals that recognize the same recurrent epitope to determine if the antibody sequences are truly identical or highly similar. Because we have identified the recurrent epitopes, we can use them to sort for B cells that express membrane-bound B cell receptors that recognize the epitope. We would do this for B cells from multiple individuals and sequence the sorted B cell's antibody gene. Identical sequences across individuals would suggest that at least a portion of the antibodies generated by humoral immune system are not completely random as generally believed. It is possible that the inherent biases in antibody gene rearrangement were evolutionarily selected to ensure production of antibodies that recognize epitopes on common viruses.

Knowledge of these highly recurrent epitopes could be used to create a more focused version of VirScan optimized for rapidly diagnosing viral infections. In the paper, we showed that on average approximately 1% of a virus's peptides are recurrent and that the recurrent peptides alone can actually increase the sensitivity and specificity of diagnosing certain viral infections. Thus, for applications that focus on diagnosis, it may be possible to assay 100-fold fewer peptides without sacrificing performance.

With such few peptides, it should be possible to use a peptide microarray instead of bacteriophage display for the assay. Peptide microarrays contain peptides synthesized or immobilized on specific locations on a solid substrate (*116*). Serum antibodies are allowed to bind to the peptides and a secondary antibody is used to visualize which peptides the serum antibodies bind. The advantage of peptide microarrays is the assay time is much faster than with high-throughput sequencing, which may be important in a clinical setting. Using a secondary antibody for IgM antibodies can distinguish between acute infection and long-term antibodies, since IgM antibodies are generally only found in the early stages of the adaptive immune response.

It may also be possible to use the knowledge of these highly recurrent epitopes to enhance efficacy of vaccines. In particular, we believe these epitopes can be fused to subunit vaccines to elicit immune complex formation *in vivo*, which should increase the potency of the immune response.

Immune complexes are composed of antibodies bound to soluble antigens. They can enhance the humoral immune response by activating the complement cascade (*117*), promoting Fc-receptor mediated endocytosis and antigen processing by antigen presenting cells (*118*), and

enabling Fc-receptor mediated presentation of antigens on follicular dendritic cells for selection and affinity maturation of B cells (*119*).

Vaccination with pre-formed immune complexes is known to enhance primary and secondary antibody responses to model antigens as well as viruses such as hepatitis B virus and human immunodeficiency virus (*120*, *121*). However, production of immune complex vaccines is difficult because it requires formulating an antigen in complex with an antibody that binds it.

In our paper, we reported highly recurrent epitopes from very common human viruses. Over 90-95% of the samples we studied had antibodies against these epitopes. If these highly recurrent epitopes are fused to an antigen for vaccination, the fused antigen should form immune complexes *in vivo* with pre-existing antibodies against the recurrent epitope. With this method, no exogenous antibodies are required because it "piggybacks" on the antibodies that are already present in the vast majority of the population. In fact, except for adding a short sequence to the gene encoding the antigen, the rest of the vaccine manufacturing process can be left unchanged, allowing production to occur essentially using currently established methods and facilities. This simple modification should be compatible with standard manufacturing procedures and could significantly improve the efficacy of current and future vaccines.

Mechanistic studies of scleroderma subclasses

We discovered a novel subclass of patients with scleroderma who have autoantibodies against the minor spliceosome complex. Using the complementary approaches PhIP-Seq and PLATO, we detected autoantibodies against many, but not all of the components of the complex. It would be interesting to use other approaches, perhaps ELISA or IP-Western with proteins expressed in human cells, to determine the extent of epitope spreading within this complex.

Knowing the extent of autoimmunity against the minor spliceosome complex will also inform experiments to determine this subclass's association with malignancy. Approximately half of the patients in this subclass had coincident diagnoses of cancer and scleroderma. A previous study showed that in patients with autoantibodies against Pol III, the autoantibodies are a result of an immune response against a tumor-associated mutant POLR3A gene that cross-reacts with the wild-type protein (*26*).

To look for evidence of a similar mechanism operating in the newly discovered subclass, we could take tumor biopsies from the patients with coincident diagnoses of cancer and scleroderma and sequence the genes encoding components of the minor spliceosome complex. As a control, we could sequence the same genes in patients with coincident diagnoses but no autoantibodies against the minor spliceosome complex. If genetic alterations in these genes were only found in patients with autoantibodies against the minor spliceosome, it would suggest that the alterations might trigger a cross-reactive response. Because the mutant sequence is known, it would then be possible to look for mutant-specific immune responses that may have preceded autoimmunity.

Knowledge of this novel subclass could also be used for clinical purposes. Our results indicate that autoantibodies against the RNPC3 gene product is found specifically in patients with scleroderma and not in healthy people or patients with other autoimmune diseases. Thus, this autoantibody specificity could be a useful diagnostic biomarker for scleroderma, which can be difficult to diagnose. In addition, the three previously characterized subclasses are known to be associated with different disease severities (*97*, *99–101*). Studies of the onset and progression

of disease in the novel subclass may also reveal that this autoantibody specificity would be a useful prognostic marker as well.

The study of adaptive immunity needs high-throughput tools

The incredible flexibility and specificity of the antibody response is dependent on the highly diverse repertoire of naïve and affinity-matured antibodies. Utilizing and understanding the power of this response requires high-throughput approaches. In addition, the immune system is exposed to a wide variety of environmental antigens and a host of ever-evolving microorganisms. Although model antigens and inbred organisms are powerful tools, native immune responses are much more complex. The combined diversity of possible antibody-antigen combinations present formidable challenges for study, but, as described in the projects of this thesis, new techniques that take advantage of the rapid advances in DNA sequencing and synthesis can enable investigations of this complexity at unprecedented depth.

In addition to purely DNA sequencing-based studies of the host immune response, creating large collections of antigens of interest using DNA synthesis enables a much more complete picture of the immune system in the context of antigen exposure. The tools used and described in this thesis – PhIP-Seq, PLATO, and VirScan – provide complementary approaches to tackling the diversity of the antibody response to viral and autoantigens.

Beyond the humoral response, the adaptive immune response also has a cellular component mediated by T cells. Similar high-throughput approaches could be developed to study these responses, but they will require the development of a novel assay format because T cell receptors are always membrane bound, are generally lower affinity, and recognize their antigen in the context of the major histocompatibility protein on the surface of an antigen presenting cell.

The combination of such a method with the ones described in this thesis will enable a much greater understanding of the complex interplay within the adaptive immune response in infection and autoimmunity.

Appendix A:

Recombination of barcoded libraries

In this appendix section, I describe the design and implementation of a strategy for the recombination of two barcoded DNA libraries that enables unique identification of the recombined pairs of DNA using high-throughput short read sequencing of the adjacent barcodes.

In this dissertation, I have described several applications of the power of high-throughput sequencing for identifying members of a population that are enriched in selection experiments. This approach works very well for constructs encoded by short nucleic acid sequences, since most high-throughput sequencing technologies have a relatively short maximum read length and pursuing longer read lengths comes at the cost of the number of reads, limiting the complexity of the libraries which can be screened. Thus for larger constructs, such as full length ORFs, our laboratory has begun using barcoded DNA libraries, in which the ORFs are paired with a short DNA barcode during cloning. These barcodes uniquely identify the ORF, thus overcoming the limitations of the sequencing technologies. In addition, amplifying just the barcodes instead of full length ORFs limits the introduction of PCR biases.

This approach requires an initial paired-end sequencing step to map which barcodes correspond to which embers of the library (ORFs, in this case). Because the ORFs have very different sequences, even short reads which cover only a fragment of the ORF can uniquely identify it. However, this approach would not work if the members of the library shared significant sequence similarity, for example with a paired ORF library, where one ORF could be paired with many other ORFs, or a recombinant antibody library in which the one heavy chain sequence could be paired with many other light chain sequences. This is because the sequencing technologies have a limit on the maximum size of the amplicon. So the paired end sequencing would only be able to read the barcode and the proximal DNA fragment, but that isn't enough for

unique identification because the same proximal DNA fragment can pair with multiple different distal DNA fragments. One strategy would be to introduce restriction sites flanking the proximal DNA fragment such that cutting and intramolecular ligation would bring the barcode and the distal DNA fragment closer, but this is a laborious process that may introduce biases and would need to be performed before each sequencing step. Instead, we chose to develop a library construction strategy that utilizes recombination of two separately barcoded libraries which results in adjacent barcodes identifying the constructs from both libraries so that a simple PCR will capture both adjacent barcodes which uniquely identify the members of both libraries.

The strategy builds on the mating-assisted genetically integrated cloning (MAGIC) *in vivo* cloning technology previously developed in our laboratory (*122*). Briefly, MAGIC uses the *E. coli* transfer system of F factor (F') to conjugate a donor plasmid with a conditional origin of replication (R6K oriγ) and an insert flanked by 50 bp homology arms and a rare cleavage site (for I-SceI) from a permissive host to a non-permissive host that expresses the endonuclease I-SceI and the lambda-Red recombination system with a recipient plasmid containing the same homology arms flanking the rare cutter sites. The I-SceI cuts the donor and recipient plasmids and the lambda-Red recombination system recombines the insert fragment into the recipient plasmid. Appropriate selectable markers on the insert and donor DNA allow selection for the recombined product.

MAGIC is very efficient, but to increase it's efficiency even further, I substituted M13 phage instead of conjugation to transfer the donor plasmid to the recipient strain. This required the introduction of the phage f1 origin or replication into the donor plasmid and packaging using a standard M13 helper phage. I also placed the barcodes for the insert and recipient DNA directly

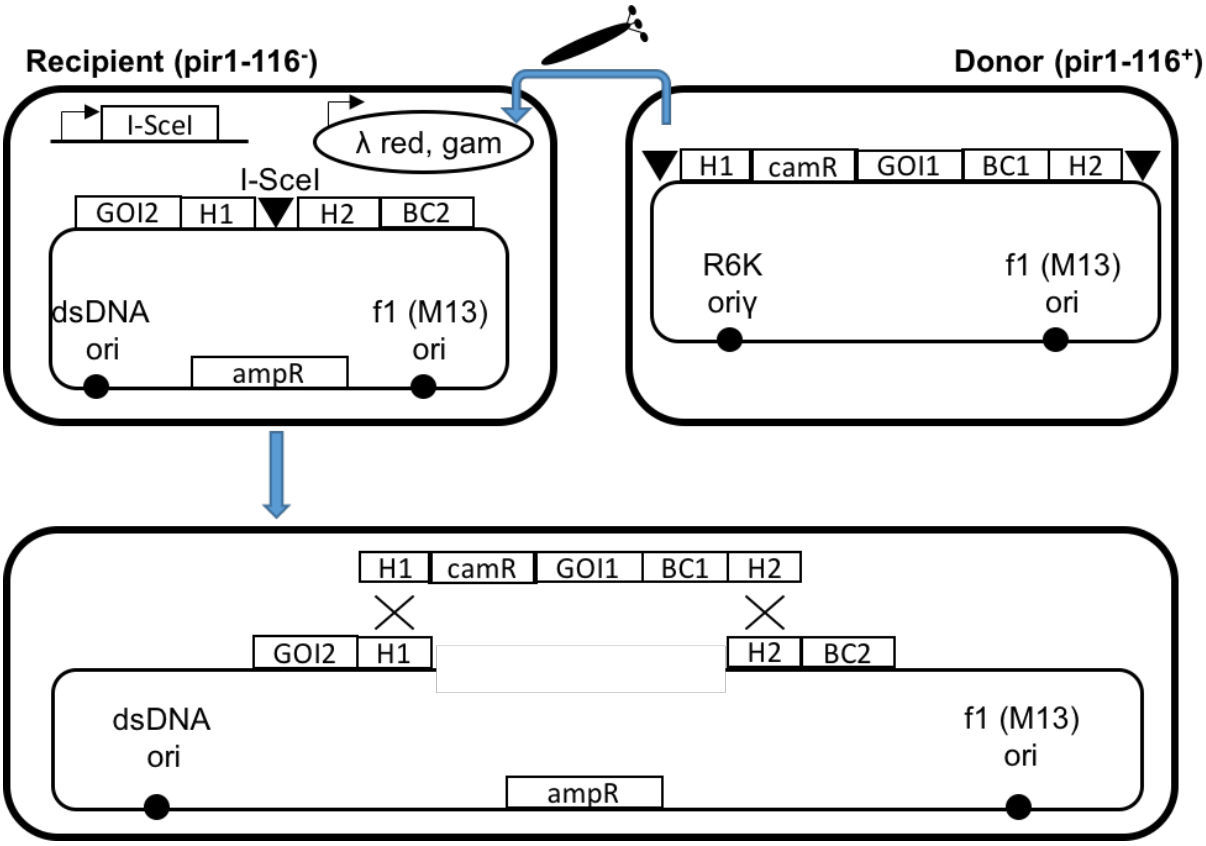

Figure 33. Schematic for barcoded DNA library recombination strategy. The donor *E. coli* strain (top right) expresses a relaxed copy-number control allele of the *trans*-acting factor π encoded by the *pir1-116*. This factor is necessary for the replication of the donor plasmid which is under the control of R6K ori γ. The donor plasmid contains a chloramphenicol resistance gene (camR), followed by a gene of interest (GOI1) and a barcode (BC1). This sequence is flanked by two 50-bp homology arms (H1, H2) and I-SceI sites (inverted black triangle). Finally the donor plasmid also contains an f1 ori for packaging in M13 phage. This phage is used to infect the recipient strain (top left). The recipient strain is deficient in pir1-116 but expresses I-SceI and the lambda red recombination proteins under inducible control. The recipient plasmid contains a single I-SceI site flanked by the same two homology arms (H1, H2) and a second gene of interest (GOI2). Following infection and induction of I-SceI and the lambda red proteins (bottom), the insert from the donor plasmid will recombine into recipient plasmid, bringing the barcodes together and the genes of interest together into one vector.

adjacent to the homology arms. Finally, I included a selectable marker (chloramphenicol resistance gene) and a separate restriction site in the insert DNA so I could easily assay for successful recombination (Figure 33).

After infection and induction of I-SceI and the lambda Red gene, I plated the cells on amipicillin (selectable marker for the recipient plasmid) and obtained approximately 10^8 cfu/mL, consistent with the optical density of the cultures (see Methods). Then, I replica-plated 48 colonies onto plates containing ampicillin and chloramphenicol to assay for the presence of the insert DNA. I repeated this procedure three times and obtained between 12 and 19 ampicillin- and chloramphenicol-resistant colonies, suggesting that the recombination efficiency is between 25-40%. In one replicate, I also picked 10 of the resistant colonies to miniprep and performed a double digest using restriction enzymes that cut once each in the recipient and donor plasmid. In all 10 colonies, the double digest resulted in the banding pattern predicted for the recombined product (Figure 34). Finally, I sequenced the DNA from these 10 colonies, which confirmed that all 10 were properly recombined.

These results indicate that the recombination is highly efficient. In addition, substituting M13 phage for conjugation greatly improved efficiency. The original MAGIC paper demonstrated approximately $2*10^6$ cfu/mL recombinants, whereas with M13 phage, I was able to obtain $2.5\text{-}4.0*10^7$ cfu/mL, an increase of over an order of magnitude. This strategy will be useful for constructing barcoded DNA libraries for sequencing-assisted selections.

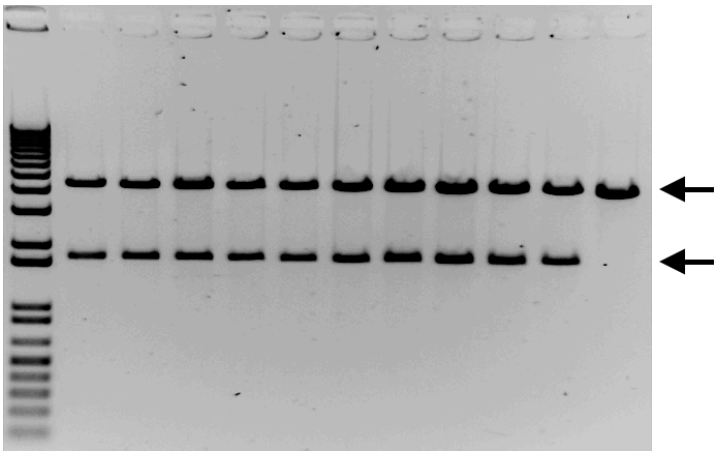

Figure 34. Restriction digest of recombinants. The first lane is a molecular weight ladder. The next ten are individual ampicillin and chloramphenicol resistant colonies following double-digest. The last lane is the recipient plasmid following double digest. The arrows indicate the predicted size of the restriction digest products.

Methods

To prepare the donor phage, the donor strain BUN20 (*122*) containing the donor plasmid was grown at 30 ºC in Luria-Bertani broth containing 12.5 µg/mL of chloramphenicol to A_{600}=0.3. M13K07 helper phage was added at an MOI of 20:1 and the culture was allowed to incubate at 30 ºC for another 30 min. 50 µg/mL of kanamycin was added to the culture and grown overnight. The overnight culture was centrifuged at 10,000 *g* for 10 min and the supernatant was supplemented with one volume of 2.5 M NaCl/20% PEG-8000 for each four volumes of supernatant. The mixture was allowed to sit for 30 min at 4 ºC and then centrifuged for 10,000 *g* for 10 min and the precipitant was resuspended in 1/10 the volume of PBS. This mixture of M13 phage was then titered using the standard protocol.

To perform the recombination, the donor strain BUN21 containing the recombination plasmid pML104 and the recipient plasmid (*122*) was grown overnight at 30 ºC in Luria-Bertani

supplemented with 100 μg/mL of ampicillin, 50 μg/mL of spectinomycin, and 0.2% w/v of glucose. Following overnight growth, the culture was washed twice in two volumes of Luria-Bertani broth. The recipient strain was then diluted 1:200 in Luria-Bertani broth and grown to A_{600}=0.15-0.25, at which point the donor phage was added at an MOI of 20:1. The media was supplemented with 0.2% w/v L-arabinose and the culture was incubated at 37 ºC for 2 h without shaking followed by 2 h with shaking. Serial dilutions of the cultures were grown overnight at 42 ºC on selection plates containing 100 μg/mL of ampicillin.

Works Cited

1. C. A. Janeway, The immune system evolved to discriminate infectious nonself from noninfectious self. *Immunol. Today*. **13**, 11–16 (1992).

2. C. A. Janeway, R. Medzhitov, Innate immune recognition. *Annu. Rev. Immunol.* **20**, 197–216 (2002).

3. C. H. Bassing, W. Swat, F. W. Alt, The mechanism and regulation of chromosomal V(D)J recombination. *Cell*. **109** (2002), , doi:10.1016/S0092-8674(02)00675-X.

4. A. I. Apostoaei, J. R. Trabalka, Review, Synthesis, and Application of Information on the Human Lymphatic System to Radiation Dosimetry for Chronic Lymphocytic Leukemia (2010).

5. H. W. Schroeder Jr., Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev Comp Immunol*. **30**, 119–135 (2006).

6. D. R. Davies, E. A. Padlan, S. Sheriff, Antibody-Antigen Complexes. *Annu. Rev. Biochem.* **59**, 439–473 (1990).

7. G. Georgiou *et al.*, The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32**, 158–68 (2014).

8. P. A. Carr, G. M. Church, Genome engineering. *Nat. Biotechnol.* **27**, 1151–1162 (2009).

9. A. Phillipidis, The Top 25 Best-Selling Drugs of 2014. *Genet. Eng. Biotechnol. News* (2015), (available at http://www.genengnews.com/insight-and-intelligenceand153/the-top-25-best-selling-drugs-of-2014/77900383/).

10. D. C. Andersen, D. E. Reilly, Production technologies for monoclonal antibodies and their fragments. *Curr. Opin. Biotechnol.* **15** (2004), pp. 456–462.

11. M. Uhlen *et al.*, Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28** (2010), pp. 1248–1250.

12. K. Jonasson, L. Berglund, M. Uhlen, The 6th HUPO Antibody Initiative (HAI) workshop: Sharing data about affinity reagents and other recent developments. *Proteomics*. **10** (2010), pp. 2066–2068.

13. H. B. Larman, G. Jing Xu, N. N. Pavlova, S. J. Elledge, Construction of a rationally designed antibody platform for sequencing-assisted selection. *Proc. Natl. Acad. Sci.* **109**, 18523–18528 (2012).

14. M. Fasnacht *et al.*, Automated antibody structure prediction using Accelrys tools: Results and best practices. *Proteins Struct. Funct. Bioinforma.* **82**, 1583–1598 (2014).

15. A. Sivasubramanian, A. Sircar, S. Chaudhury, J. J. Gray, Toward high-resolution homology modeling of antibody F v regions and application to antibody-antigen docking. *Proteins Struct. Funct. Bioinforma.* **74**, 497–514 (2009).

16. M. Pedotti, L. Simonelli, E. Livoti, L. Varani, Computational docking of antibody-antigen complexes, opportunities and pitfalls illustrated by influenza hemagglutinin. *Int. J. Mol. Sci.* **12**, 226–251 (2011).

17. F. Delunardo *et al.*, Screening of a microvascular endothelial cDNA library identifies rabaptin 5 as a novel autoantigen in Alzheimer's disease. *J. Neuroimmunol.* **192**, 105–112 (2007).

18. M. B. Soares *et al.*, Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 9228–9232 (1994).

19. H. B. Larman *et al.*, Autoantigen discovery with a synthetic human peptidome. *Nat. Biotechnol.* **29**, 535–541 (2011).

20. S. Kosuri, G. M. Church, Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods*. **11**, 499–507 (2014).

21. E. Hammarlund *et al.*, Duration of antiviral immunity after smallpox vaccination. *Nat. Med.* **9**, 1131–1137 (2003).

22. J. L. Mokili, F. Rohwer, B. E. Dutilh, Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* **2**, 63–77 (2012).

23. J. Zhu *et al.*, Protein interaction discovery using parallel analysis of translated ORFs (PLATO). *Nat. Biotechnol.* **31**, 331–4 (2013).

24. T. R. Katsumoto, M. L. Whitfield, M. K. Connolly, The pathogenesis of systemic sclerosis. *Annu. Rev. Pathol.* **6**, 509–537 (2011).

25. S. I. Nihtyanova, C. P. Denton, Autoantibodies as predictive tools in systemic sclerosis. *Nat. Rev. Rheumatol.* **6**, 112–116 (2010).

26. C. G. Joseph *et al.*, Association of the autoimmune disease scleroderma with an immunologic response to cancer. *Science*. **343**, 152–157 (2014).

27. A. A. Shah, A. Rosen, L. Hummers, F. Wigley, L. Casciola-Rosen, Close temporal relationship between onset of cancer and scleroderma in patients with RNA polymerase I/III antibodies. *Arthritis Rheum.* **62**, 2787–2795 (2010).

28. U. Ravn *et al.*, By-passing in vitro screening - Next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res.* **38** (2010), doi:10.1093/nar/gkq789.

29. H. Zhang *et al.*, Phenotype-information-phenotype cycle for deconvolution of combinatorial antibody libraries selected against complex systems. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13456–13461 (2011).

30. C. F. Barbas, Synthetic human antibodies. *Nat. Med.* **1**, 837–839 (1995).

31. C. F. Barbas, J. D. Bain, D. M. Hoekstra, R. A. Lerner, Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 4457–4461 (1992).

32. E. Vargas-Madrazo, F. Lara-Ochoa, J. C. Almagro, Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition. *J. Mol. Biol.* **254**, 497–504 (1995).

33. C. V. Lee *et al.*, High-affinity human antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold. *J. Mol. Biol.* **340**, 1073–1093 (2004).

34. C. Lloyd *et al.*, Modelling the human immune response: performance of a 1011 human antibody repertoire against a broad panel of therapeutically relevant antigens. *Protein Eng. Des. Sel.* **22**, 159–168 (2009).

35. W. Zhai *et al.*, Synthetic antibodies designed on natural sequence landscapes. *J. Mol. Biol.* **412**, 55–71 (2011).

36. F. Ehrenmann, Q. Kaas, M.-P. Lefranc, IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.* **38**, D301–7 (2010).

37. Q. Kaas, M. Ruiz, M.-P. Lefranc, IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res.* **32**, D208–10 (2004).

38. J. Hanes, C. Schaffitzel, A. Knappik, A. Pluckthun, Picomolar affinity antibodies from a fully synthetic naive library selected and evolved by ribosome display. *Nat. Biotechnol.* **18**, 1287–1292 (2000).

39. Y. Ofran, A. Schlessinger, B. Rost, Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B cell epitopes. *J. Immunol.* **181**, 6230–6235 (2008).

40. F. Schutz, M. Delorenzi, MAMOT: hidden Markov modeling tool. *Bioinformatics*. **24**, 1399–1400 (2008).

41. C. J. Bond, C. Wiesmann, J. C. J. Marsters, S. S. Sidhu, A structure-based database of antibody variable domain diversity. *J. Mol. Biol.* **348**, 699–709 (2005).

42. F. Lara-Ochoa, E. Vargas-Madrazo, M. A. Jimenez-Montano, J. C. Almagro, Patterns in the complementary determining regions of immunoglobulins (CDRs). *Biosystems.* **32**, 1–9 (1994).

43. H. Singh, G. P. Raghava, ProPred: prediction of HLA-DR binding sites. *Bioinformatics*. **17**, 1236–1237 (2001).

44. S. Fabre-Lafay *et al.*, Nectin-4 is a new histological and serological tumor associated marker for breast cancer. *BMC Cancer*. **7**, 73 (2007).

45. A. M. Athanassiadou, E. Patsouris, A. Tsipis, M. Gonidi, P. Athanassiadou, The significance of Survivin and Nectin-4 expression in the prognosis of breast carcinoma. *Folia Histochem. Cytobiol.* **49**, 26–33 (2011).

46. M. Chodorge, L. Fourage, G. Ravot, L. Jermutus, R. Minter, In vitro DNA recombination by L-Shuffling during ribosome display affinity maturation of an anti-Fas antibody increases the population of improved variants. *Protein Eng. Des. Sel.* **21**, 343–351 (2008).

47. H. R. Hoogenboom, Selecting and screening recombinant antibody libraries. *Nat. Biotechnol.* **23**, 1105–1116 (2005).

48. A. Beck, T. Wurch, C. Bailly, N. Corvaia, Strategies and challenges for the next generation of therapeutic antibodies. *Nat. Rev. Immunol.* **10**, 345–352 (2010).

49. Y. Erlich *et al.*, DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res.* **19**, 1243–1253 (2009).

50. S. Prabhu, I. Pe'er, Overlapping pools for high-throughput targeted resequencing. *Genome Res.* **19**, 1254–1261 (2009).

51. D. R. Bowley, T. M. Jones, D. R. Burton, R. A. Lerner, Libraries against libraries for combinatorial selection of replicating antigen-antibody pairs. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 1380–1385 (2009).

52. T. A. Whitehead *et al.*, Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30**, 543–548 (2012).

53. D. J. Schofield *et al.*, Application of phage display to high throughput antibody generation and characterization. *Genome Biol.* **8**, R254 (2007).

54. P. Zacchi, D. Sblattero, F. Florian, R. Marzari, A. R. M. Bradbury, Selecting open reading frames from DNA. *Genome Res.* **13**, 980–990 (2003).

55. D. G. Gibson *et al.*, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*. **6**, 343–345 (2009).

56. J. Hanes, A. Pluckthun, In vitro selection and evolution of functional proteins by using ribosome display. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 4937–4942 (1997).

57. C. Zahnd, P. Amstutz, A. Pluckthun, Ribosome display: selecting and evolving proteins in vitro that specifically bind to a target. *Nat. Methods*. **4**, 269–279 (2007).

58. J. Hanes, L. Jermutus, S. Weber-Bornhauser, H. R. Bosshard, A. Pluckthun, Ribosome display efficiently selects and evolves high-affinity antibodies in vitro from immune libraries. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14130–14135 (1998).

59. M. J. Feldhaus *et al.*, Flow-cytometric isolation of human antibodies from a nonimmune Saccharomyces cerevisiae surface display library. *Nat. Biotechnol.* **21**, 163–170 (2003).

60. C. Schaffitzel, J. Hanes, L. Jermutus, A. Pluckthun, Ribosome display: an in vitro method for selection and evolution of antibodies from libraries. *J. Immunol. Methods*. **231**, 119–135 (1999).

61. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. **17**, 10 (2011).

62. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 1–10 (2009).

63. K. M. Wylie, G. M. Weinstock, G. A. Storch, Emerging view of the human virome. *Transl. Res.* **160**, 283–290 (2012).

64. B. a Duerkop, L. V Hooper, Resident viruses and their interactions with the immune system. *Nat. Immunol.* **14**, 654–659 (2013).

65. E. S. Barton *et al.*, Herpesvirus latency confers symbiotic protection from bacterial infection. *Nature*. **447**, 326–329 (2007).

66. E. F. Foxman, A. Iwasaki, Genome-virome interactions: examining the role of common viral infections in complex disease. *Nat. Rev. Microbiol.* **9**, 254–264 (2011).

67. M. Lecuit, M. Eloit, The human virome: New tools and concepts. *Trends Microbiol.* **21**, 510–515 (2013).

68. I. De Vlaminck *et al.*, Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell*. **155**, 1178–1187 (2013).

69. The UniProt Consortium, Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42**, D191–198 (2014).

70. H. B. Larman *et al.*, PhIP-Seq characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis. *J. Autoimmun.* **43**, 1–9 (2013).

71. C. Bialecki, H. M. Feder, J. M. Grant-Kels, The six classic childhood exanthems: a review and update. *J. Am. Acad. Dermatol.* **21**, 891–903 (1989).

72. J. H. Lee, W. K. Roth, S. Zeuzem, Evaluation and comparison of different hepatitis C virus genotyping and serotyping assays. *J. Hepatol.* **26**, 1001–1009 (1997).

73. H. F. L. Wertheim *et al.*, Key role for clumping factor B in Staphylococcus aureus nasal colonization of humans. *PLoS Med.* **5**, 0104–0112 (2008).

74. R. A. Manz, A. E. Hauser, F. Hiepe, A. Radbruch, Maintenance of serum antibody levels. *Annu. Rev. Immunol.* **23**, 367–386 (2005).

75. M. Wang *et al.*, Human anti-JC virus serum reacts with native but not denatured JC virus major capsid protein VP1. *J. Virol. Methods*. **78**, 171–176 (1999).

76. S. A. S. Staras *et al.*, Seroprevalence of cytomegalovirus infection in the United States, 1988-1994. *Clin. Infect. Dis.* **43**, 1143–1151 (2006).

77. M. A. Reynolds, D. Kruszon-Moran, A. Jumaan, D. S. Schmid, G. M. McQuillan, Varicella seroprevalence in the U.S.: data from the National Health and Nutrition Examination Survey, 1999-2004. *Public Health Rep.* **125**, 860–869.

78. J. I. Cohen, Epstein–Barr virus infection. *N. Engl. J. Med.* **343**, 481–492 (2000).

79. L. Dong *et al.*, A combination of serological assays to detect human antibodies to the avian influenza a H7N9 virus. *PLoS One*. **9**, e95612 (2014).

80. P. Patel *et al.*, Prevalence and Risk Factors Associated With Herpes Simplex Virus-2 Infection in a Contemporary Cohort of HIV-Infected Persons in the United States. *Sex. Transm. Dis.* **39**, 154–160 (2012).

81. C. T. Stover *et al.*, Prevalence of and risk factors for viral infections among human immunodeficiency virus (HIV)-infected and high-risk HIV-uninfected women. *J. Infect. Dis.* **187**, 1388–1396 (2003).

82. E. A. Engels *et al.*, Risk factors for human herpesvirus 8 infection among adults in the United States and evidence for sexual transmission. *J. Infect. Dis.* **196**, 199–207 (2007).

83. R. Vita *et al.*, The Immune Epitope Database 2.0. *Nucleic Acids Res.* **38**, D854–862 (2009).

84. H. Singh, H. R. Ansari, G. P. S. Raghava, Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence. *PLoS One*. **8**, e62216 (2013).

85. J. Zhu *et al.*, Protein interaction discovery using parallel analysis of translated ORFs (PLATO). *Nat. Biotechnol.* **31**, 331–334 (2013).

86. Y. Urwijitaroon, S. Teawpatanataworn, A. Kitjareontarm, Prevalence of cytomegalovirus antibody in Thai-northeastern blood donors. *Southeast Asian J. Trop. Med. Public Health*. **24 Suppl 1**, 180–182 (1993).

87. M. J. Cannon, D. S. Schmid, T. B. Hyde, Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. *Rev. Med. Virol.* **20**, 202–213 (2010).

88. S. Mohanna *et al.*, Human herpesvirus-8 in Peruvian blood donors: a population with hyperendemic disease? *Clin. Infect. Dis.* **44**, 558–561 (2007).

89. D. Ablashi *et al.*, Seroprevalence of human herpesvirus-8 (HHV-8) in countries of Southeast Asia compared to the USA, the Caribbean and Africa. *Br. J. Cancer*. **81**, 893–897 (1999).

90. J. S. Smith, N. J. Robinson, Age-specific prevalence of infection with herpes simplex virus types 2 and 1: a global review. *J. Infect. Dis.* **186 Suppl** , S3–S28 (2002).

91. A. Heit *et al.*, CpG-DNA aided cross-priming by cross-presenting B cells. *J. Immunol.* **172**, 1501–1507 (2004).

92. Y. Aydar, S. Sukumar, A. K. Szakal, J. G. Tew, The influence of immune complex-bearing follicular dendritic cells on the IgM response, Ig class switching, and production of high affinity IgG. *J. Immunol.* **174**, 5358–5366 (2005).

93. M. F. Quigley *et al.*, Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 19414–19419 (2010).

94. K. J. L. Jackson, M. J. Kidd, Y. Wang, A. M. Collins, The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. *Front. Immunol.* **4**, 263 (2013).

95. P. Parameswaran *et al.*, Convergent antibody signatures in human dengue. *Cell Host Microbe*. **13**, 691–700 (2013).

96. M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **5** (2010), doi:10.1101/pdb.prot5448.

97. A. Gabrielli, E. V Avvedimento, T. Krieg, Scleroderma. *N. Engl. J. Med.* **360**, 1989–2003 (2009).

98. C. Muangchan *et al.*, The 15% Rule in Scleroderma: The Frequency of Severe Organ Complications in Systemic Sclerosis. A Systematic Review. *J. Rheumatol.* **40** (2013), pp. 1545–1556.

99. M. Koenig, M. Dieudé, J. L. Senécal, Predictive value of antinuclear autoantibodies: The lessons of the systemic sclerosis autoantibodies. *Autoimmun. Rev.* **7** (2008), pp. 588–593.

100. Y. Okano, V. D. Steen, T. A. Medsger, Autoantibody reactive with RNA polymerase III in systemic sclerosis. *Ann. Intern. Med.* **119**, 1005–1013 (1993).

101. J. C. Parker, R. W. Burlingame, T. T. Webb, C. C. Bunn, Anti-RNA polymerase III antibodies in patients with systemic sclerosis detected by indirect immunofluorescence and ELISA. *Rheumatology*. **47**, 976–979 (2008).

102. M. J. Mamula, Epitope spreading: The role of self peptides and autoantigen processing by B lymphocytes. *Immunol. Rev.* **164**, 231–239 (1998).

103. H. Benecke, R. Lührmann, C. L. Will, The U11/U12 snRNP 65K protein acts as a molecular bridge, binding the U12 snRNA and U11-59K protein. *EMBO J.* **24**, 3057–3069 (2005).

104. C. Netter, G. Weber, H. Benecke, M. C. Wahl, Functional stabilization of an RNA recognition motif by a noncanonical N-terminal expansion. *RNA*. **15**, 1305–1313 (2009).

105. J. J. Turunen, E. H. Niemelä, B. Verma, M. J. Frilander, The significant other: Splicing by the minor spliceosome. *Wiley Interdiscip. Rev. RNA*. **4** (2013), pp. 61–76.

106. R. J. Smeenk, Antinuclear antibodies: cause of disease or caused by disease? *Rheumatology (Oxford)*. **39** (2000), pp. 581–584.

107. Z. Y. Chen *et al.*, Immune complexes and antinuclear, antinucleolar, and anticentromere antibodies in scleroderma. *J. Am. Acad. Dermatol.* **11**, 461–467 (1984).

108. S. Jordan *et al.*, Effects and safety of rituximab in systemic sclerosis: an analysis from the European Scleroderma Trial and Research (EUSTAR) group. *Ann. Rheum. Dis.*, 1–7 (2014).

109. C. G. Joseph *et al.*, Association of the autoimmune disease scleroderma with an immunologic response to cancer. *Science*. **343**, 152–7 (2014).

110. X. Zhou *et al.*, HLA-DPB1 and DPB2 are genetic loci for systemic sclerosis: A genome-wide association study in Koreans with replication in North Americans. *Arthritis Rheum.* **60**, 3807–3814 (2009).

111. H. B. Larman, A. C. Liang, S. J. Elledge, J. Zhu, Discovery of protein interactions using parallel analysis of translated ORFs (PLATO). *Nat. Protoc.* **9**, 90–103 (2014).

112. A. R. M. Bradbury, S. Sidhu, S. Dübel, J. McCafferty, Beyond natural antibodies: the power of in vitro display technologies. *Nat. Biotechnol.* **29**, 245–254 (2011).

113. R. M. Horton, H. D. Hunt, S. N. Ho, J. K. Pullen, L. R. Pease, Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. *Gene*. **77**, 61–68 (1989).

114. S. S. Sidhu *et al.*, Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. *J. Mol. Biol.* **338**, 299–310 (2004).

115. M. Gebauer, A. Skerra, Engineered protein scaffolds as next-generation antibody therapeutics. *Curr. Opin. Chem. Biol.* **13** (2009), pp. 245–255.

116. M. Beyer *et al.*, Combinatorial synthesis of peptide arrays onto a microchip. *Science*. **318**, 1888 (2007).

117. G. G. Klaus, The generation of memory cells. II. Generation of B memory cells with preformed antigen-antibody complexes. *Immunology*. **34**, 643–652 (1978).

118. B. C. Schalke, W. E. Klinkert, H. Wekerle, D. S. Dwyer, Enhanced activation of a T cell line specific for acetylcholine receptor (AChR) by using anti-AChR monoclonal antibodies plus receptors. *J. Immunol.* **134**, 3643–3648 (1985).

119. B. Heyman, The immune complex: possible ways of regulating the antibody response. *Immunol. Today*. **11**, 310–313 (1990).

120. E. Celis, T. Chang, Antibodies to hepatitis B surface antigen potentiate the response of human T lymphocyte clones to the same antigen. *Science (80-.).* **224**, 297–299 (1984).

121. U. Abdel-Motal, S. Wang, S. Lu, K. Wigglesworth, U. Galili, Increased Immunogenicity of Human Immunodeficiency Virus gp120 Engineered To Express Gal 1-3Gal 1-4GlcNAc-R Epitopes. *J. Virol.* **80**, 6943–6951 (2006).

122. M. Z. Li, S. J. Elledge, MAGIC, an in vivo genetic method for the rapid construction of recombinant DNA molecules. *Nat. Genet.* **37**, 311–319 (2005).